VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

Bayesian Asymptotics Under Misspecification

Kleijn, B.J.K.

2004

Link to publication in VU Research Portal

citation for published version (APA)
Kleijn, B. J. K. (2004). *Bayesian Asymptotics Under Misspecification*. [PhD-Thesis - Research and graduation internal, Vrije Universiteit Amsterdam].

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 05. May. 2023

Bayesian Asymptotics Under Misspecification

B.J.K. Kleijn

Bayesian Asymptotics Under Misspecification

Kleijn, Bastiaan Jan Korneel

Bayesian Asymptotics Under Misspecification/ B.J.K. Kleijn.
-Amsterdam : Vrije Universiteit Amsterdam, Faculteit der Exacte Wetenschappen
Proefschrift Vrije Universiteit Amsterdam. -Met lit. opg.
-Met samenvatting in het Nederlands.
ISBN 90-9017838-4
Trefw.: Bayesiaanse statistiek, Asymptotische statistiek; NUR 916, 918

Bayesian Asymptotics Under Misspecification

Bayesiaanse Asymptotiek Onder Misspecificatie

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Vrije Universiteit Amsterdam, op gezag van de rector magnificus prof. dr. T. Sminia, in het openbaar te verdedigen ten overstaan van de promotiecommissie van de Faculteit der Exacte Wetenschappen op vrijdag 19 maart 2003 om 10.45 uur in de Aula van de universiteit, De Boelelaan 1105

door

Bastiaan Jan Korneel Kleijn

geboren op 15 oktober 1970, te Nijmegen.

Promotor: Prof. dr. A.W. van der Vaart

Faculteit der Exacte Wetenschappen
Afdeling Wiskunde
Vrije Universiteit Amsterdam

Het onderzoek besproken in dit proefschift is grotendeels gefinancierd door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek, middels een project getiteld "Statistical Estimation in the Errors in Variables Model", projectnummer 613.003.043. Het proefschrift zelf is tot stand gekomen met financiele steun van het Thomas Stieltjes Instituut voor Wiskundig Onderzoek.

NWO

THOMAS STIELTJES INSTITUTE
FOR MATHEMATICS

Contents

Chapter 1

Introduction

Motivation

Many problems in statistics are estimation problems for independent and identically distributed measurements: given a set $\mathscr{P}$ of candidate probability distributions (the model) and a given set of n measurement results $X_1, \ldots, X_n$ (the sample) distributed *i.i.d.* according to an unknown distribution P_0, give an estimate $\hat{P}_n$ in $\mathscr{P}$ for P_0. Sometimes the goal is more modest and one is interested only in certain characteristic properties of the underlying distribution P_0. Simple examples are the mean or median of X, more complicated are the degree of smoothness or the number of maxima of the density of P_0. These estimation problems usually involve certain assumptions about P_0, if only to make the questions meaningful: if we consider properties of the density for X, we have already assumed that it is distributed continuously; if we are interested in its mean, the implied assumption is that of integrability. Other reasons to place conditions on the true distribution are often motivated from a mathematical perspective: some theorems only hold under certain, often quite natural conditions on P_0.

Obviously, estimation problems require a *choice* for the model $\mathscr{P}$. Usually, the model choice is made on the basis of interpretability of the parameterisation, in relation to the specific quantities that play a role in the background of the problem. For instance, the ever-popular normal model (which comprises all normal distributions on the space in which measurements take their values) leaves mean and variance of the data to be estimated, both of which are readily interpreted. In less extreme cases, (parametric or non-parametric) models are chosen as reasonable approximations to the well-specified situation (although it is often left unclear in which sense the approximation is reasonable). Another reason for the use of simple but misspecified models is, again, mathematical convenience: a small collection of candidate distributions often makes the estimation problem less complicated. Note that generically we can not guarantee that P_0 lies in the model beforehand, so the choice of a model amounts

to a bias in the estimate. This bias will be smaller if a larger model is chosen. (Whether or not such a bias is problematic depends on the context of the estimation problem.) So there is a trade-off here: on the one hand, a small, restrictive model leads to interpretability and mathematical simplicity, on the other hand, a large model leads to answers that are closer to the truth.

These two aspects of statistical estimation, the choice for the model $\mathscr{P}$ and assumptions on the underlying distribution P_0 are linked by the assumption that the model is *well specified*, *i.e.*

$$P_0 \in \mathscr{P}. \tag{1.1}$$

Properties of P_0 are then implied by the choice of $\mathscr{P}$. Again, mathematical convenience often dictates this assumption: the mathematical analysis is greatly simplified if we know that P_0 is among the candidates beforehand. In fact, this assumption is so common in mathematical statistics that it is omitted in the statement of theorems habitually. In applied statistics, theorems that rely on (1.1) are often used without mention of the fact that, in all likelihood, the true distribution of the data does not lie in the model.

However, there is a good reason for applied statisticians to 'abuse' these theorems in this way: they often work regardless! To the mathematical statistician this raises the question why, *i.e.* "Is it possible to prove those same theorems *without* the assumption that $P_0 \in \mathscr{P}$?". That summarises exactly the point of view we adopt in this thesis. Note that it is not implied that there will be no conditions on P_0: the point is to formulate conditions on P_0 that are weaker than (1.1) above. The resulting restrictions on P_0 delimit the theorem's range of applicability more appropriately (if not entirely). We shall speak of a misspecified model in that case. In principle, theoretical results for misspecified models give the statistician more freedom in his choice of model, because interpretability and well specification cease to compete while the accuracy of approximation (the bias mentioned earlier) plays only a minimal role.

The theorems that we consider from the misspecified point of view concern the asymptotic behaviour of Bayesian procedures. Asymptotic statistics asks the question what happens to statistical procedures in the large-sample limit. Bayesian statistics can be viewed as the strict consequence of interpreting the likelihood as an (unnormalised) probability density, as suggested already by the word 'likelihood'. As it turns out, theorems on Bayesian consistency, rate of convergence and limiting shape of the posterior distribution can be proved using a suitably defined point of convergence within a (possibly misspecified) model (be it the true distribution P_0 or some 'closest' alternative P^*). The right definition for the closest alternative turns out to be the point in the model at minimal Kullback-Leibler divergence with respect to P_0, while other conditions remain comparable to those found in the well-specified situation.

Given the range of application of present-day statistics and more particularly, Bayesian statistics, this should be useful beyond the extent of a mere academic exercise in (applied) mathematics.

The current chapter provides an introduction to Bayesian statistics, asymptotic statistics,

Bayesian asymptotics and model misspecification. It has been written for the broadest possible audience (which implies that for some the presentation may be too elementary) and has two primary goals: firstly to discuss the most important concepts used in following chapters at an introductory level and secondly to formulate appropriate analogies between Bayesian asymptotic concepts and similar concepts in the theory of point estimation. Most of this chapter is concerned with well-specified, smooth, parametric estimation problems. This way, readers unfamiliar with Bayesian methods are offered the opportunity to acquaint themselves with the most important definitions and their properties in a relatively simple (and well-specified) context, before meeting the same concepts in more complicated situations in later chapters. For instance, in section 1.3 we consider asymptotic consistency in Bayesian statistics, formulating it as weak convergence of the posterior to a degenerate measure located at the point in the (well-specified) model that corresponds to the true distribution. We also indicate how this definition relates to consistency as defined in point estimation. The analogous definition of 'consistency' in misspecified situations will be given in chapter 2.

The next three chapters each contain an article on a specific aspect of the asymptotic behaviour of Bayesian methods in situations where the model is misspecified. Each is preceded by a brief introduction relating it to the rest of the thesis.

1.1 Bayesian statistics

In this section, we consider the basic definitions of Bayesian statistics with the emphasis on the definition of prior and posterior distributions. Furthermore, we discuss some aspects of Bayesian point estimation and the choice of the prior distribution. The discussion as presented here is necessarily very brief. Various books providing an overview of Bayesian statistics can be recommended, depending on the background and interest of the reader: a very theoretical treatment can be found in Le Cam (1986) [67]. For a more mundane version, the reader is referred to Van der Vaart (1998) [91] and Le Cam and Yang (1990) [68]. A general and fairly comprehensive reference of a more practical inclination is Berger (1985) [6] and finally, Ripley (1996) [79] discusses matters with (decision-theoretical, pattern-classification) applications in mind but does not loose sight of the statistical background.

1.1.1 Prior and posterior distributions

Formalising the Bayesian procedure can be done in several ways. We start this subsection with considerations that are traditionally qualified as being of a 'subjectivist' nature, but eventually we revert to the 'frequentist' point of view. Concretely this means that we derive an expression for the posterior and prove regularity in the subjectivist framework. In a frequentist setting, this expression is simply used as a definition and properties like regularity and measurability are imposed. Ultimately, the philosophical motivation becomes irrelevant from the mathematical point of view once the posterior and its properties are established.

The observation Y lies in a space $\mathscr{Y}$ with σ-algebra $\mathscr{B}$ and the model Θ is assumed to be a measurable space as well, with σ-algebra $\mathscr{G}$. Perhaps the most elegant (and decidedly subjectivist) Bayesian framework unifies observation space and model as follows: we start from the product-space $\mathscr{Y} \times \Theta$ with product σ-algebra $\mathscr{F} = \sigma(\mathscr{B} \times \mathscr{G})$ and probability measure $\underline{\Pi} : \sigma(\mathscr{B} \times \mathscr{G}) \to [0,1]$ which is *not* a product measure. The marginal probability Π on $\mathscr{G}$ is called the prior and is interpreted as the subjectivist's 'degree of belief' attached to subsets of the model *a priori* (that is, before any observation has been made). The fact that we have defined a probability measure on the product of sample space and model makes it possible to condition on Y or on $\underline{\theta}$ (in particular). The conditional probability distribution[1] $\Pi_{Y|\underline{\theta}} : \mathscr{B} \times \Theta \to [0,1]$ is such that:

(i) for every $A \in \mathscr{B}$, the map $\theta \mapsto \Pi_{Y|\underline{\theta}}(A, \theta)$ is $\mathscr{G}$-measurable,

(ii) for Π-almost-all $\theta \in \Theta$, the map $A \mapsto \Pi_{Y|\underline{\theta}}(A, \theta)$ defines a probability measure.

The measures $\Pi_{Y|\underline{\theta}}(\,\cdot\,|\underline{\theta} = \theta)$ form a (Π-almost-sure) version of the elements P_θ of the model $\mathscr{P}$:

$$P_\theta = \Pi_{Y|\underline{\theta}}(\,\cdot\,|\,\underline{\theta} = \theta\,) : \mathscr{B} \to [0,1]$$

Consequently, frequentist's notion of a model can only be represented up to null-sets of the prior in this setting.

Specific to the Bayesian framework is the conditional probability distribution:

$$\Pi_{\underline{\theta}|Y} : \mathscr{G} \times \mathscr{Y} \to [0,1], \tag{1.2}$$

which is called the posterior distribution and can be interpreted as a version of the prior corrected by observation of Y through conditioning. If we choose $\mathscr{Y}$ equal to the n-fold product of the sample space $\mathscr{X}$ (with σ-algebra $\mathscr{A}$), the observation is written as $Y = (X_1, X_2, \ldots, X_n)$. The additional assumption that the sample is *i.i.d.* (a statement concerning the *conditional* independence of the observations given $\underline{\theta} = \theta$) takes the form:

$$\Pi_{Y|\underline{\theta}}(\,X_1 \in A_1, \ldots, X_n \in A_n \,|\, \underline{\theta} = \theta\,) = \prod_{i=1}^{n} \Pi_{Y|\underline{\theta}}(\,X_i \in A_i \,|\, \underline{\theta} = \theta\,) = \prod_{i=1}^{n} P_\theta(X_i \in A_i),$$

[1] The precise definition of the conditional distribution is rather subtle: we may define the sub-σ-algebra $\mathscr{C} = \{\mathscr{Y} \times G : G \in \mathscr{G}\}$ and condition by means of the $\underline{\Pi}$-almost-sure definition $\underline{\Pi}_{Y|\mathscr{C}}(A, \omega) = E[1_A|\mathscr{C}](\omega)$ ($\omega = (y, \theta) \in \mathscr{Y} \times \Theta$) for each $A \in \mathscr{B}$ separately. However, the fact that exceptional $\underline{\Pi}$-null-sets depend on A may render the map $\underline{\Pi}_{Y|\mathscr{C}} : \mathscr{B} \times (\mathscr{Y} \times \Theta) \to [0,1]$ ill-defined, since its domain of definition has an exceptional set equal to the union of exceptional null-sets over all $A \in \mathscr{B}$. We require that $\underline{\Pi}_{Y|\mathscr{C}}(\,\cdot\,, \omega)$ is well-defined $\underline{\Pi}$-almost-surely as a probability measure, that is, as a *map* on all of $\mathscr{B}$ rather that for each A separately (in which case, $\underline{\Pi}_{Y|\mathscr{C}}$ is called a regular conditional probability). A sufficient condition for the existence of a regular version of $\underline{\Pi}_{Y|\mathscr{C}}$ is that Y is a so-called *Polish space*, *i.e.* a complete, separable, metric space with Borel σ-algebra (see *e.g.* Dudley (1989) [28], section 10.2 and in particular theorem 10.2.2). Note also that due to the special choice for $\mathscr{C}$, $\mathscr{C}$-measurability implies that $\underline{\Pi}_{Y|\mathscr{C}}(\,\cdot\,, (y, \theta))$ depends on θ alone. Hence we denote it $\Pi_{Y|\underline{\theta}} : \mathscr{B} \times \Theta \to [0,1]$.

for all $(A_1, \ldots, A_n) \in \mathscr{A}^n$. Assuming that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure on $\mathscr{X}$, the above can also be expressed in terms of μ-densities $p_\theta : \mathscr{X} \to \mathbb{R}$. Using Bayes' rule, we obtain the following expression for the posterior distribution:

$$\Pi_n(\, G \,|\, X_1, X_2, \ldots, X_n \,) = \frac{\displaystyle\int_G \prod_{i=1}^n p(X_i)\, d\Pi(P)}{\displaystyle\int_{\mathscr{P}} \prod_{i=1}^n p(X_i)\, d\Pi(P)}, \tag{1.3}$$

where $G \in \mathscr{G}$ is a measurable subset of the model $\mathscr{P}$. Note that we have simplified our notation for the posterior somewhat (compare with (1.2)) and omitted representation in terms of the variable θ. Finally, there exists also a marginal for the observations which takes the form:

$$\Pi_Y(\, X_1 \in A_1, \ldots, X_n \in A_n \,) = \int_\Theta \prod_{i=1}^n P_\theta(A_i)\, d\Pi(\theta)$$

This distribution is called the prior predictive distribution and describes a Bayesian's expectations concerning observations $X_1, X_2, \ldots, X_n$ based on the prior Π. Note the relation with de Finetti's theorem, which says that the distribution of a sequence $(X_1, \ldots, X_n)$ of random variables is of the form on the *r.h.s.* of the above display (with uniquely determined prior Π) if and only if the $X_1, \ldots, X_n$ are exchangeable. The prior predictive distribution is subject to correction by observation through substitution of the prior by the posterior: the resulting posterior predictive distribution is interpreted as the Bayesian's expectation concerning the distribution of the observations $X_{n+1}, X_{n+2}, \ldots$ given the observations $X_1, X_2, \ldots, X_n$.

We conclude this discussion of the distributions that play a role in Bayesian statistics with an important point: note that at no stage an 'underlying distribution of the sample' was used or needed. For this reason, the 'pure' Bayesian is reluctant to assume the existence of a distribution P_0 for the sample.

The distribution P_0 could not play a role in this thesis if we did not choose to adopt a different, rather more frequentist point of view: we assume the sample of observations in $\mathscr{X}$ to be *i.i.d.* P_0-distributed and we have a model $\mathscr{P}$ which is a probability space $(\mathscr{P}, \mathscr{G}, \Pi)$ with a probability measure Π which we refer to as the prior. In this way, model and sample space are left in the separate roles they are assigned by the frequentist. We then proceed to *define* the posterior by expression (1.3). To guarantee measurability, we assume (or impose) that the properties (i) and (ii) above are satisfied (see Schervish (1995) [81] and Barron, Schervish and Wasserman (1999) [5] for a detailed analysis).

Finally, we note that it is not necessary that the model is dominated. One easily shows that posterior can be rewritten using the Radon-Nikodym derivative[2] of P with respect to

[2]The measure P can be decomposed uniquely in a P_0-absolutely-continuous part $P_\parallel$ and a P_0-singular part $P_\perp$: $P = P_\parallel + P_\perp$. Following Le Cam, we use the convention that if P is not dominated by P_0, the Radon-Nikodym derivative refers to the P_0-absolutely-continuous part only: $dP/dP_0 = dP_\parallel/dP_0$.

P_0:

$$\Pi_n(\, A \,|\, X_1, X_2, \ldots, X_n \,) = \frac{\displaystyle\int_A \prod_{i=1}^{n} \frac{dP}{dP_0}(X_i)\, d\Pi(P)}{\displaystyle\int_{\mathscr{P}} \prod_{i=1}^{n} \frac{dP}{dP_0}(X_i)\, d\Pi(P)}, \quad (P_0^n - a.s.) \tag{1.4}$$

In cases where the model is not dominated, (1.4) may be used as the definition of the posterior measure. Alternatively, any σ-finite measure that dominates P_0 may be used instead of P_0 in (1.4) while keeping the definition P_0^n-almost-sure. This is used in chapter 3.

1.1.2 Bayesian point estimation

The link between Bayesian procedures and ordinary (point-)estimation methods is provided by estimator sequences derived from the posterior. The most straightforward, if rather primitive way is by a simply drawing a point from the posterior distribution. However, there exist other methods that do not randomise once the posterior is determined: to give a few examples, we consider a model $\mathscr{P}$ with metric d. First of all, we define the posterior mean (or posterior expectation):

$$\hat{P}_n = \int_{\mathscr{P}} P \; d\Pi_n(\, P \,|\, X_1, \ldots, X_n \,), \tag{1.5}$$

if P is suitably integrable with respect to $\Pi_n(\,\cdot\,|\, X_1, \ldots, X_n\,)$. Note that unless $\mathscr{P}$ is convex, $\hat{P}_n \in \mathscr{P}$ is not guaranteed. Also note that if we consider a measurable map $\theta \mapsto P_\theta$ of a (convex) parameter-set Θ with prior measure $\Pi(d\theta)$ onto a space of probability measures $\mathscr{P}$ (with induced prior $\Pi(dP)$), it makes a difference whether we consider the posterior mean as defined in (1.5), or calculate $P_{\hat{\theta}_n}$. More generally, we may define so-called *formal Bayes estimators* [67] as minimisers over the model of functions:

$$P \mapsto \int_{\mathscr{P}} \ell_n(d(P,Q))\, d\Pi_n(\, Q \,|\, X_1, \ldots, X_n \,),$$

where ℓ_n is a sequence of convex loss functions. If the model and the map $P \mapsto d^2(P,Q)$ are convex, the posterior mean can be viewed as a formal Bayes estimator if we choose $\ell_n(x) = x^2$; the posterior median is obtained if we choose $\ell_n(x) = |x|$. Another useful point estimator derived from the posterior is defined (for given $\epsilon > 0$) as a maximiser of the function:

$$P \mapsto \Pi_n(\, B_d(P,\epsilon) \,|\, X_1, \ldots, X_n \,),$$

where $B_d(P,\epsilon)$ is the d-ball in $\mathscr{P}$ of radius ϵ centred on P. Similarly, for fixed p such that $1/2 < p < 1$, we may define a point estimator by the centre point of the smallest d-ball with posterior mass greater than or equal to p. If the posterior is dominated by a σ-finite measure μ, we can define the so-called *maximum a posteriori* estimator as a maximum of the posterior density (sometimes referred to as the posterior mode). Note that a different choice of dominating measure leads to a different definition of the MAP estimator.

1.1.3 The choice of the prior

Bayesian procedures have been the object of much criticism, often focusing on the choice of the prior as an undesirable source of ambiguity. The answer of the subjectivist that the prior represents the 'belief' of the statistician or 'expert knowledge' pertaining to the measurement elevates this ambiguity to a matter of principle, thus setting the stage for a heated debate between 'pure' Bayesians and 'pure' Frequentists concerning the philosophical merits of either school within statistics. The issue is complicated further by the fact that the basic setup for the Bayesian procedure, as described in subsection 1.1.1, does not refer to the 'true' distribution P_0 for the observation, providing another point of fundamental philosophical disagreement for the fanatically pure to lock horns over.

Leaving the philosophical argumentation to others, we shall try to discuss the choice of a prior at a more practical level. First, let us seek the appropriate analogy with the choices that are made in ordinary point estimation. The support of the prior[3] can be viewed as the Bayesian analog of the choice of model in frequentist statistics. If the true, underlying distribution lies outside the model (or the support of the prior), one can speak of misspecification, so this point is of considerable importance for the following chapters.

The subjective aspect of a prior is more subtle because it depends not only on the support but on all details of the prior distribution: even when the support of the prior is fixed, there is a large collection of possible priors left to be considered, each leading to a different posterior distribution. Arguably, the freedom of choice left in ordinary point estimation at this stage concerns all possible point estimators within the chosen model, so the choice of a prior is not the embarrassment of riches it is sometimes made out to be.

Subjectiveness finds its mathematical expression when high prior belief or expert knowledge are translated in relatively large amounts of assigned prior mass for certain regions of the model. Attempts to minimise the amount of such prejudice introduced by the prior therefore focus on uniformity (argumentation that departs from the Shannon entropy in discrete probability spaces reaches the same conclusion (see, for example, Ghosh and Ramamoorthi (2003) [41], p. 47)). The original references on Bayesian methods (Bayes (1763) [3], Laplace (1774) [61]) use uniform priors as well. Uniformity of the prior introduces a number of complications, however. If we consider, for example, a parametric model $\Theta \subset \mathbb{R}^k$ and $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$, straightforward uniformity of the prior on Θ would force us to choose it proportional to the Lebesgue measure. However, we would have to require that Θ be compact in order to maintain our definition of the prior as a probability measure. Even if this rather restrictive condition is accepted, new problems with this construction emerge. The sense in which uniformity is achieved here is parameterisation-dependent since uniformity on Θ has no intrinsic meaning for $\mathscr{P}$. For twice-differentiable parametric models, a construction from Riemannian geometry can be borrowed to define a parameterisation-independent prior (see Jeffreys (1946), (1961)

[3] If the σ-algebra on the model is the Borel-σ-algebra, the support of a measure is defined as the set of all points for which all open neighbourhoods have measure strictly greater than zero.

[50, 51]) if we interpret the Fisher information as a Riemannian metric on the model (as first proposed by Rao (1945) [77] and extended by Efron (1975) [29]; for an overview, see Amari (1990) [1]) and use the square-root of its determinant as a density with respect to the Lebesgue measure:

$$\Pi(d\theta) = \sqrt{\det(I_\theta)}\, d\theta.$$

Other constructions and criteria exist; for an overview of these non-informative priors (also referred to as non-subjective, reference (see Bernardo (1979) [11], Berger and Bernardo (1992) [7]) or objective priors), the reader is referred to Kass and Wasserman (1995) [53]. Returning to definition (1.3), we see that it is not strictly necessary to have a probability measure Π to define the posterior. Multiplication of Π by an arbitrary non-zero constant leaves the posterior invariant and, in fact, it is possible to have an infinite measure Π (a so-called *improper* prior), as long as we can somehow guarantee that the posterior remains a probability measure. Non-informative priors, including Jeffreys prior in many situations, are often improper. Finally, certain classes of probability measures are closed under the operation of conditioning on the sample: so if we choose a prior in such a class, then the whole sequence of posteriors lies in that class as well. Such classes are called conjugate classes (and one speaks of a conjugate prior).

Fortunately, subjectiveness of the prior turns out to be a concern in finite-sample statistics primarily; for the asymptotic properties of the Bayesian procedure the important aspects of the choice of prior are of an entirely different nature as we shall see in section 1.3.

Bayesian procedures can be implemented using Markov Chain Monte Carlo simulation. The simplicity of MCMC algorithms, their applicability in non-parametric situations and the fact that they generate samples from the posterior that can be used directly to approximate integrals (*e.g.* the posterior mean), give Bayesian methods a very broad computational range. The interested reader is referred to the overview article by Escobar and West (1995) [30] and the comprehensive book by Robert (2001) [78].

1.2 Asymptotic statistics

Given an infinite *i.i.d.* sample $X_1, X_2, \ldots$ drawn from P_0 and a model $\mathscr{P}$, an estimation procedure prescribes a sequence of estimates $\hat{P}_n \in \mathscr{P}$, each calculated using only the first n observations. More generally, any statistical procedure can be indexed by the size n of the sample used to calculate it, leading to sequences of (parameter) estimates, tests, confidence regions, *etcetera*. Properties of such sequences reflect the behaviour of the estimation procedure with growing sample-size. An intuitively reasonable requirement of any estimation procedure is a property known as consistency: the sequence $\hat{P}_n$ approaches the true distribution P_0 to within arbitrary precision if the sample on which the estimation is based is made large enough. Similarly, samples of arbitrarily large size should enable one to test with power arbitrarily close to one and define arbitrarily small confidence regions. Further analysis of a

consistent sequence $\hat{P}_n$ concerns the (suitably rescaled) distribution of the estimator-sequence around its point of convergence. The mathematical formulation of these concepts is based on so-called limit theorems, which describe the behaviour of an estimation procedure in the limit that the number of measurements goes to infinity.

The study of the asymptotic regime of an estimation procedure is interesting for two reasons. First of all, asymptotic results provide approximations to exact values; finite-sample calculations often become intractable, even in relatively simple (parametric) situations. However, the analysis of the large-sample limit is often still possible when finite-sample procedures are intractable or otherwise hard to carry out exactly. The answer obtained in the large-sample limit may then be used as an approximation[4] to the finite-sample answer (asymptotic confidence intervals are a good example). Secondly, if we have several possible estimation procedures available for a certain problem, asymptotic behaviour provides us with the means to compare their performance on (very) large samples. Of course, the first performance criterion is consistency. To choose between two consistent procedures, one may consider rate of convergence and properties of the limit distribution characterising the degree of concentration (like asymptotic variance or asymptotic risk).

In this section, we give an overview of the aspects of asymptotic point estimation that are most important for the following chapters. It should be noted that this discussion is not intended to be comprehensive, nor is it stretched to full generality: the focus is on the asymptotic regime of smooth parametric estimation methods and even in that respect we are far from complete. For a more comprehensive presentation, the reader is referred to some of the excellent textbooks on the subject, like Ibragimov and Has'minskii (1981) [47], Le Cam and Yang (1990) [68] and Van der Vaart (1998) [91].

1.2.1 Consistency, rate and limit distribution

A sequence of estimators $\hat{P}_n$ is said to be (asymptotically) consistent (respectively almost-surely consistent) if the estimator converges to the true distribution P_0 in probability (respectively almost-surely) as the sample-size goes to infinity. More precisely, a sequence $\hat{P}_n$ of estimators in a model $\mathscr{P}$ (with metric d) for $P_0 \in \mathscr{P}$ is said to be consistent if:

$$d(\hat{P}_n, P_0) \xrightarrow{P_0} 0.$$

In the case of a parametric model (with k-dimensional parameter set Θ, open in $\mathbb{R}^k$) defined by $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ with metric $d(P_{\theta_1}, P_{\theta_2}) = \|\theta_1 - \theta_2\|$, estimation of θ_0 (such that $P_0 = P_{\theta_0}$) by $\hat{\theta}_n$ is consistent if $\hat{\theta}_n \xrightarrow{P_0} \theta_0$. A consistent estimator may be analysed further by appraisal

[4] A valid objection to the use of asymptotic approximations is the fact that this practice does not provide any relation between the accuracy of the approximation and the size n of the sample for which answers are approximated. Limit theorems guarantee that approximation errors fall below arbitrarily small bounds for 'large enough' n, but do not specify what 'large enough' is exactly. It is common practice to ignore this fact and assume that the asymptotic answer is a 'good' approximation for sample sizes that are deemed 'large enough'.

of its rate of convergence and limit distribution. Let us define the rate of convergence first: a sequence r_n such that

$$r_n^{-1}\, d(\hat{P}_n, P_0) = O_{P_0}(1), \tag{1.6}$$

i.e. the *l.h.s.* in the above display is bounded in probability[5], then r_n is an upper bound to the rate of convergence of the estimator sequence $\hat{P}_n$ with respect to the metric d. The rate thus gives the speed with which balls around the point of convergence may be shrunk while still capturing the estimator with high probability. Note that Prohorov's theorem guarantees weak convergence of a subsequence of the sequence $r_n\, d(\hat{P}_n, P_0)$. In particular, the distribution of the rescaled metric distance of the estimator to the point of convergence is asymptotically tight along the subsequence. Heightening the level of detail even further, we can require that the sequence of estimators, when centred on its point of convergence and rescaled by the rate, converges weakly to a non-degenerate distribution over the (localised) model, the so-called limit distribution. It should be noted that, in general, both rate and limit distribution depend not only on the quantity that is to be estimated and the used estimation procedure, but on the model and on specific properties of the point of convergence as well[6]. Specialising again to the parametric case, we say that $\hat{\theta}_n$ converges to θ_0 at rate r_n^{-1} with limit distribution L_{θ_0} if:

$$r_n^{-1}(\hat{\theta}_n - \theta_0) \overset{\theta_0}{\rightsquigarrow} L_{\theta_0}. \tag{1.7}$$

To illustrate these definitions, we consider the application of the strong law of large numbers and the central limit theorem. The strong law can be used to prove (almost-sure) consistency and the central limit theorem refines the analysis, proving $\sqrt{n}$-rate of convergence and a normal limit distribution. In parametric models with parameters that can be identified with expectations of certain integrable random variables, the law of large numbers proves consistency of *i.i.d.*-sample averages as follows. Assume that $\theta \in \Theta$ parameterises the model and that there is a $\theta_0 \in \Theta$ such that $P_{\theta_0} = P_0$. Let the parameter θ_0 for the true distribution equal the expectation $P_{\theta_0} T$ for some ($\sigma(X)$-measurable) and integrable random variable T (for example the observation X, estimating the location μ of a normal distribution $N(\mu, \sigma^2)$), the sample-average $\mathbb{P}_n T$ is a consistent estimator for θ_0 since:

$$\mathbb{P}_n T = \frac{1}{n} \sum_{i=1}^{n} T(X_i) \overset{P_{\theta_0}-a.s.}{\longrightarrow} P_{\theta_0} T = \theta_0,$$

which implies convergence in P_{θ_0}-probability in particular. If, in addition, the random variable T is square-integrable, the sample-average converges at rate $1/\sqrt{n}$ with a normal limit distribution due to the central limit theorem:

$$\mathbb{G}_n T = \sqrt{n}(\mathbb{P}_n - P_{\theta_0}) T \overset{P_{\theta_0}}{\rightsquigarrow} N(0, \mathrm{Var}_{P_{\theta_0}}(T)).$$

---

[5] Written out in full, this amounts to the requirement that for every $\epsilon > 0$ there exists a bound $M > 0$ such that $\sup\{P_0^n(r_n^{-1}\, d(\hat{P}_n, P_0) > M) : n \geq 1\} < \epsilon$.

[6] More accurately, the rate depends on the *size* of the model (see *e.g.* chapter 3 and section 7 in Ghosal and Van der Vaart (2001) [40]). The dependence of the limit distribution on model and point of convergence is the subject of subsection 1.2.2, which is closely related to the material presented in chapter 2.

Using quantiles of the normal distribution on the *r.h.s.* the above provides an asymptotic approximation of confidence regions as alluded to in the introduction of this section. Together with the delta-rule[7] (see Van der Vaart (1998) [91], chapter 3) and the maximum-likelihood estimator (see the end of this subsection), the above completes the basic 'asymptotic toolkit' for parametric models. Its remarkable simplicity is also its strength and the range of applications for the above is endless.

Other estimation methods often rely on the strong law and central limit theorem as well, albeit indirectly: for instance, the moment method solves (for some integrable f, originally $f(x) = x^k$ for some $k \geq 1$, whence the name 'moment method') the equation $\mathbb{P}_n f = P_\theta f$ for θ. The strong law asserts that $\mathbb{P}_n f$ converges to $P_{\theta_0} f$, P_{θ_0}-almost-surely. If the dependence $\theta \mapsto e(\theta) = P_\theta f$ is one-to-one in an open neighbourhood of the point of convergence, the moment-estimator takes the form: $\hat{\theta}_n = e^{-1}(\mathbb{P}_n f)$ for large enough n. If, in addition, f is square-integrable with respect to P_{θ_0} and e^{-1} is differentiable with non-singular derivative at θ_0, the delta-rule guarantees that the moment estimator $\hat{\theta}_n$ is asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}\big(e^{-1}(\mathbb{P}_n f) - e^{-1}(P_{\theta_0} f)\big) \overset{\theta_0}{\rightsquigarrow} N(0, \Sigma), \tag{1.8}$$

with covariance matrix equal to $\Sigma = (e^{-1})'(e_0)^T P_{\theta_0} f f^T (e^{-1})'(e_0)$, where $e_0 = P_{\theta_0} f$.

Another important class of estimators consists of so-called M-estimators, which are defined as maximisers of criterion functions $M_n : \Theta \to \mathbb{R}$ (dependent on $(X_1, \ldots, X_n)$) with respect to $\theta \in \Theta$: the estimators $\hat{\theta}_n$ satisfy[8]

$$M_n(\hat{\theta}_n) = \sup_{\theta \in \Theta} M_n(\theta),$$

(assuming, of course, that the supremum exists in $\mathbb{R}$). Often the criterion function depends on the sample through the empirical distribution: $M_n(\theta) = \mathbb{P}_n m_\theta$, where m_θ is a $\sigma(X)$-measurable random variable. If m_θ is P_0-integrable for every θ, point-wise convergence of the form:

$$M_n(\theta) = \mathbb{P}_n m_\theta \overset{P_0-a.s.}{\longrightarrow} P_0 m_\theta = M(\theta), \tag{1.9}$$

is guaranteed by the strong law. Given that we choose the random variable m_θ such that the maximum of M over Θ exists, is unique and well-separated, it would seem reasonable to expect (near-)maximisers $\hat{\theta}_n$ of M_n to converge to the maximiser of M. A suitable choice for m guarantees that a maximiser of $M(\theta)$ coincides with θ_0 such that $P_{\theta_0} = P_0$. Note, however, that the condition of maximisation is a global one, whereas (1.9) provides only point-wise convergence. So the strong law may be insufficient to prove consistency of M-estimators. Indeed a basic proof of consistency can be given under the rather strong condition that the

[7] The delta-rule roughly says that the application of differentiable functions to estimator sequences leaves the rate unchanged (if the derivative does not happen to be zero in θ_0) and induces a linear transformation on the limiting random variable, see *e.g.* (1.8).

[8] In many circumstances, the definition of an M-estimator sequence can be relaxed to include *near*-maximisers of the criterion function, which satisfy $M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta) - o_{P_0}(a_n)$ for some sequence $a_n \downarrow 0$ to be specified (see the formulation of theorem 1.1 for an example).

class $\{m_\theta : \theta \in \Theta\}$ is Glivenko-Cantelli, (although this requirement is far stronger than is actually needed). A well-known alternative set of conditions for consistency of M-estimation was given by Wald[9] [94]. For a more elaborate discussion of consistent M-estimation in parametric models, see *e.g.* section 5.2 in Van der Vaart (1998) [91].

Once consistency is established, M-estimators are asymptotically normal under differentiability conditions and a Lipschitz condition on the function m_θ, (which together constitute a set of so-called *regularity* conditions, see the next subsection) as exemplified in the following theorem. We assume that Θ is open in $\mathbb{R}^k$, that $P_0 = P_{\theta_0}$ for the point θ_0 in Θ that maximises $\theta \mapsto P_0 m_\theta$, and that $\hat{\theta}_n$ is a consistent estimator sequence for θ_0, defined as near-maximisers of criterion functions as in (1.9).

Theorem 1.1. *For each $\theta \in \Theta$, let $x \mapsto m_\theta(x)$ be a measurable function such that $\theta \mapsto m_\theta(X)$ is P_0-almost-surely differentiable at θ_0 with derivative $\dot{m}_{\theta_0}(X)$. Furthermore, suppose that there exists a P_0-square-integrable random variable $\dot{m}$ such that for all θ_1, θ_2 in a neighbourhood of θ_0:*

$$\left| m_{\theta_1}(X) - m_{\theta_2}(X) \right| \leq \dot{m}(X) \, \|\theta_1 - \theta_2\|, \quad (P_0 - a.s.). \tag{1.10}$$

Let the map $\theta \mapsto P_0 m_\theta$ have a second-order Taylor expansion around θ_0:

$$P_0 m_\theta = P_0 m_{\theta_0} + \tfrac{1}{2}(\theta - \theta_0)^T V_{\theta_0}(\theta - \theta_0) + o(\|\theta - \theta_0\|^2), \quad (\theta \to \theta_0). \tag{1.11}$$

with non-singular second-derivative matrix V_{θ_0}. Then any consistent sequence of estimators $\hat{\theta}_n$ such that $\mathbb{P}_n m_{\hat{\theta}_n} \geq \sup_\theta \mathbb{P}_n m_\theta - o_{P_0}(n^{-1})$ satisfies:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_{\theta_0}^{-1} \dot{m}_{\theta_0}(X_i) + o_{P_0}(1).$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $V_{\theta_0}^{-1} \, P_0 \left[\dot{m}_{\theta_0} \dot{m}_{\theta_0}^T \right] V_{\theta_0}^{-1}$.

Proof The proof of this theorem can be found in Van der Vaart (1998) [91], p. 54. □

Note that the Taylor-expansion (1.11) lacks a first-order term because θ_0 maximises $\theta \mapsto P_0 m_\theta$ and $\partial_\theta [P_0 m_\theta]_{\theta=\theta_0} = P_0 \partial_\theta [m_\theta]_{\theta=\theta_0}$ as a result of the domination condition (1.10).

We mention one M-estimation procedure in particular: the maximum-likelihood estimator, which maximises the criterion function $\theta \mapsto \mathbb{P}_n \log p_\theta$ with p_θ the density of P_θ with respect to a suitable dominating measure for the model[10]. If the map $\theta \mapsto \log p_\theta(X)$ is P_{θ_0}-almost-surely differentiable for all θ, we define the so-called score-function $\dot{\ell}_\theta$ as the vector of partial derivatives of the log-likelihood at θ:

$$\dot{\ell}_\theta(X) = \partial_\theta [\log p_\theta(X)]. \tag{1.12}$$

[9] Wald's conditions are actually fairly close to the Glivenko-Cantelli property, see section 5 in Van der Vaart (1999) [92] for a detailed explanation)

[10] The condition of domination for the entire model may be weakened: note that the criterion function is well-defined P_0^n-almost-surely if we use $p_\theta = dP_\theta / dP_0$, even though P_0 may not dominate all P_θ.

Then the point $\theta = \hat{\theta}_n$ of maximum-likelihood solves the so-called score-equation:

$$\mathbb{P}_n \dot{\ell}_\theta = 0,$$

Score-equations are sometimes easier to solve than the maximisation problem for the log-likelihood. If the function $\theta \mapsto \log p_\theta$ satisfies the conditions of theorem 1.1 and the second-order Taylor coefficient V_{θ_0} equals[11] the Fisher information $I_{\theta_0} = P_0 \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}$ at the point θ_0, then the maximum-likelihood estimator $\hat{\theta}_n$ converges to θ_0 as follows:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X_i) + o_{P_0}(1). \tag{1.13}$$

Since $P_0 \dot{\ell}_{\theta_0} = 0$, the *r.h.s.* of the above display equals $I_{\theta_0}^{-1} \mathbb{G}_n \dot{\ell}_{\theta_0}$ up to order $o_{P_0}(1)$, which implies that $\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{P_0}{\rightsquigarrow} N(0, I_{\theta_0}^{-1})$ by the central limit theorem. Hence the rate of convergence equals $1/\sqrt{n}$ and the normal limit distribution has the inverse Fisher information as its covariance. Under these circumstances, the maximum-likelihood estimator is optimal, by which we mean that the rate can not be improved upon and the asymptotic variance is minimal (a property usually called (asymptotic) efficiency of estimation). To prove optimality, however, requires a far more detailed analysis which is discussed in the following subsection.

1.2.2 Local asymptotic normality

Theorem 1.1 and other theorems like it are formulated with the help of (sometimes very extensive) differentiability assumptions, supplemented with domination conditions to justify the interchange of differentiation and integration (like (1.10)). As noted by Le Cam (1970) [64], there exist very simple and well-behaved (read, asymptotically normal, parametric) examples in which such elaborate systems of conditions are not satisfied, suggesting that there exist less stringent conditions leading to the same or similar assertions. The crucial property turns out to be that of differentiability in quadratic mean, which is intimately related to a model-property called local asymptotic normality. Indeed, the 'local structure of the model' turns out to be far more important than appears to be the case from 'classical' theorems like (1.1). For many applications, knowledge of the asymptotic behaviour under the law P_{θ_0} alone is not enough and it is necessary to consider the behaviour under laws that are 'asymptotically close' to P_{θ_0} as well. For example, an unambiguous definition of optimality of estimation can only be given if we limit the class of estimators in a suitable way. The class in question consists of regular estimators and its definition depends crucially on the local structure of the model as alluded to above.

Again, we consider a parametric model Θ which we assume to be well specified, *i.e.* $P_0 = P_{\theta_0}$ for some $\theta_0 \in \Theta$, and we assume that θ_0 is an internal point of Θ. We reparameterise

[11] If $\theta \mapsto \log p_\theta$ is twice-differentiable ($P_0 - a.s.$) and differentiation and P_0-expectation may be interchanged, then $V_{\theta_0} = -I_{\theta_0}$, because $-P_0 \ddot{\ell}_{\theta_0} = P_0 \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}$, which follows from differentiation of the identity $P_0(p_\theta / p_{\theta_0}) = 1$ (valid for all θ since $P_0 = P_{\theta_0}$).

neighbourhoods of the point of convergence θ_0 in terms of a so-called local parameter h defined analogously to the *l.h.s.* of (1.7): a point $h \in \mathbb{R}^k$ corresponds to a sequence $\theta_n = \theta_0 + h/\sqrt{n}$, the tail of which falls inside Θ since θ_0 is an internal point by assumption. The first definition concerns a form of model-differentiability: a (μ-dominated) model is said to be differentiable in quadratic mean (or Hellinger differentiable) at θ_0 if there exists an $\mathbb{R}^k$-valued random vector $\dot{\ell}_{\theta_0}$ such that[12]:

$$\int \big(\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \tfrac{1}{2}h^T\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}\big)^2\,d\mu = o(\|h\|^2), \quad (h \to 0). \tag{1.14}$$

When viewed as the ($\mu - a.e.$) defining property for the vector $\dot{\ell}_{\theta_0}$, Hellinger differentiability generalises definition (1.12): if $\theta \mapsto \log p_\theta(X)$ is P_0-almost-surely differentiable at θ_0, we may choose $\dot{\ell}_{\theta_0} = \partial_\theta[\log p_\theta]_{\theta=\theta_0}$ to satisfy the above display (but the converse is not always true). The following theorem asserts that Hellinger differentiability is sufficient for an asymptotic expansion of the log-likelihood that is referred to as 'local asymptotic normality' (defined as in the assertion of the following theorem).

Theorem 1.2. *(Local asymptotic normality) Suppose that the model $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean at θ_0. Then $P_{\theta_0}\dot{\ell}_{\theta_0} = 0$ and the Fisher information matrix $I_{\theta_0} = P_{\theta_0}\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^T$ exists. Furthermore, for every converging sequence $h_n \to h$:*

$$\log\prod_{i=1}^n \frac{p_{\theta_0+h_n/\sqrt{n}}}{p_{\theta_0}}(X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T\dot{\ell}_{\theta_0}(X_i) - \frac{1}{2}h^T I_{\theta_0}h + o_{P_{\theta_0}}(1). \tag{1.15}$$

Proof A proof for this theorem can be found in van der Vaart (1998) [91], pp. 94–95. □

Note that the *r.h.s.* of the expansion (1.15) equals $\mathbb{G}_n h^T\dot{\ell}_{\theta_0}$, which converges weakly to a normal limit distribution $N(0, h^T I_{\theta_0}h)$ as a result of the central limit theorem. The second term in the expansion shifts the location of this distribution and the last term converges to zero in probability, whence we conclude that (1.15) implies weak convergence of (log-)likelihood products as follows:

$$\log\prod_{i=1}^n \frac{p_{\theta_0+h_n/\sqrt{n}}}{p_{\theta_0}}(X_i) \overset{P_{\theta_0}}{\rightsquigarrow} N(-\tfrac{1}{2}h^T I_{\theta_0}h, h^T I_{\theta_0}h).$$

The latter property of the log-likelihood in a neighbourhood of θ_0 may serve to explain the name 'local asymptotic normality', although a better explanation will be given later on in this subsection.

We have yet to specify the estimation problem: the quantity of interest is a functional $\psi : \Theta \to \mathbb{R}$ which we estimate by a sequence of statistics $T_n = T_n(X_1, \ldots, X_n)$. We assume that this can be done consistently and with $\sqrt{n}$-rate under the sequence of laws $P_{\theta_0+h/\sqrt{n}}$ for every h, so we write:

$$\sqrt{n}\Big(T_n - \psi\big(\theta_0 + \frac{h}{\sqrt{n}}\big)\Big) \overset{\theta_0+h/\sqrt{n}}{\rightsquigarrow} L_{\theta_0,h}, \tag{1.16}$$

[12] The original reference for this definition is Le Cam (1970) [64]. Many authors (see, for example [47, 85, 75, 91]) have reviewed its consequences since.

where $L_{\theta_0,h}$ is the limit distribution[13].

The following theorem (which is based on theorems 7.10 and 8.14 in Van der Vaart (1998) [91]) shows that under differentiability conditions for both model and estimated parameter, the limit of the estimators T_n in the sequence of 'localised models' $\{P_{\theta_0+h/\sqrt{n}} : h \in \mathbb{R}^k\}$ can be described in terms of a statistic in the collection of normal distributions $(L_{\theta_0,h} : h \in \mathbb{R}^k)$ (referred to as the limit experiment).

Theorem 1.3. *Assume that the model $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean at θ_0 with non-singular Fisher information matrix I_{θ_0} and that $\psi : \Theta \to \mathbb{R}$ is a functional differentiable at θ_0. Let T_n be estimators in the models $\{P_{\theta_0+h/\sqrt{n}} : h \in \mathbb{R}^k\}$ such that (1.16) holds for every h. Then there exists a randomised statistic S in the model $\{N(h, I_{\theta_0}) : h \in \mathbb{R}^k\}$ such that:*

$$\sqrt{n}\Big(T_n - \psi\big(\theta_0 + \frac{h}{\sqrt{n}}\big)\Big) \overset{\theta_0+h/\sqrt{n}}{\rightsquigarrow} S - \dot{\psi}_{\theta_0} h. \tag{1.17}$$

Proof We define:

$$\Delta_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i), \quad S_n = \sqrt{n}\big(T_n - \psi(\theta_0)\big). \tag{1.18}$$

Concerning S_n, we note that:

$$S_n = \sqrt{n}\Big(T_n - \psi\big(\theta_0 + \frac{h}{\sqrt{n}}\big)\Big) + \sqrt{n}\Big(\psi\big(\theta_0 + \frac{h}{\sqrt{n}}\big) - \psi(\theta_0)\Big).$$

The last term on the *r.h.s.* of the above display converges to $\dot{\psi}_{\theta_0} h$ as a result of the differentiability of ψ at θ_0. Hence $S_n - \dot{\psi}_{\theta_0} h$ differs from the *l.h.s.* of (1.16) by a term of order $o(1)$ and has the same weak limit. Concerning Δ_{n,θ_0}, we note that due to theorem 1.2, $P_0 \dot{\ell}_{\theta_0} = 0$ and the score-function is P_0-square-integrable, so $\Delta_{n,\theta_0} = \mathbb{G}_n \dot{\ell}_{\theta_0} \overset{P_0}{\rightsquigarrow} N(0, I_{\theta_0})$. Prohorov's theorem guarantees that the sequences Δ_{n,θ_0} and $S_n - \dot{\psi}_{\theta_0} h$ are both uniformly tight under P_0. Since marginal (uniform) tightness implies joint (uniform) tightness, we use Prohorov's theorem again (in the other direction) to conclude that the pair $(S_n - \dot{\psi}_{\theta_0} h, \Delta_{n,\theta_0})$ converges weakly along a subsequence under P_0. The marginal limits of this subsequence follow from the above: $(S_n - \dot{\psi}_{\theta_0} h, \Delta_{n,\theta_0})$ converges to (V, Δ) along a subsequence under P_0, where $V \sim L_{\theta_0,0}$ and $\Delta \sim N(0, I_{\theta_0})$. We re-define the index n, denoting the converging subsequence by $(S_n - \dot{\psi}_{\theta_0} h, \Delta_{n,\theta_0})$. Since according to (1.15), the log-likelihood converges to $h^T\Delta - \frac{1}{2} h^T I_{\theta_0} h$ in P_0-probability, we also see that

$$\Big(S_n - \dot{\psi}_{\theta_0} h, \log \prod_{i=1}^{n} \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}}(X_i)\Big) \overset{\theta_0}{\rightsquigarrow} (V, h^T\Delta - \tfrac{1}{2} h^T I_{\theta_0} h).$$

[13] In this context, it is customary to speak of the *experiment* [14] $(P_\theta : \theta \in \Theta)$ rather than the model. The distinction lies in the fact that in this case, we consider convergence of the estimators under a collection of laws in the model (*c.f.* (1.16) which holds for *all* $h \in \mathbb{R}^k$), whereas before, we used only the true distribution P_{θ_0}. Note that the estimator T_n depends only on the observations $(X_1, X_2, \ldots, X_n)$ (and not on θ directly). It can therefore be referred to as a statistic in the experiment $(P_\theta : \theta \in \Theta)$, with a law that depends on θ only through the observations. We shall not make this distinction in the following for the sake of simplicity and refer to 'experiments' as 'models' throughout.

Since $h^T\Delta \sim N(0, h^T I_{\theta_0} h)$, we find the marginal limit distribution to be normal with location and variance related as follows:

$$\log \prod_{i=1}^{n} \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}}(X_i) \overset{\theta_0}{\rightsquigarrow} N(-\tfrac{1}{2}h^T I_{\theta_0} h, h^T I_{\theta_0} h),$$

which, according to Le Cam's first lemma, implies that P_{θ_0} and the sequence $P_{\theta_0+h/\sqrt{n}}$ are mutually contiguous. We then use (the general form of) Le Cam's third lemma to conclude that:

$$S_n - \dot{\psi}_{\theta_0} h \overset{\theta_0+h/\sqrt{n}}{\rightsquigarrow} L_{\theta_0,h},$$

where the limit distribution $L_{\theta_0,h}$ is given by:

$$L_{\theta_0,h}(B) = E_{\theta_0} 1_B(V) e^{h^T\Delta - \frac{1}{2}h^T I_{\theta_0} h},$$

for all measurable B. Given (V, Δ), there exists a measurable map $V : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d \times \mathbb{R}^k$ such that for $U \sim U[0,1]$ (defined on the same probability space as (V, Δ) and independent thereof), $(V(\Delta, U), \Delta)$ and (V, Δ) have the same law (for a proof, see lemma 7.11 in Van der Vaart (1998) [91], pp. 99–100). Now, for given h, let X be distributed according to $N(h, I_{\theta_0}^{-1})$. Under $h = 0$, the distributions of $I_{\theta_0} X$ and Δ are identical and by Fubini's theorem,

$$\begin{aligned} P_h\big(V(I_{\theta_0}X, U) \in B\big) &= \int P\big((V(I_{\theta_0}x, U) \in B\big)\, dN(h, I_{\theta_0}^{-1})(x) \\ &= \int P\big((V(I_{\theta_0}x, U) \in B\big)\, \frac{dN(h, I_{\theta_0}^{-1})}{dN(0, I_{\theta_0}^{-1})}(x)\, dN(0, I_{\theta_0}^{-1})(x) = L_{\theta_0,h}(B). \end{aligned}$$

Given h, the random variable V is a randomised statistic with law $L_{\theta_0,h}$ defined in the normal model consisting of distributions of the form $N(h, I_{\theta_0}^{-1})$. Defining $S = V + \dot{\psi}_{\theta_0} h$, the weak limit of the sequence S_n given h, we establish (1.17). □

This theorem specifies the *r.h.s.* of (1.16) further in estimation problems where both the model and the functional to be estimated are smooth: the limit distribution of the estimator sequence can be viewed as the distribution of a statistic in the *normal* model $\{N(h, I_{\theta_0}) : h \in \mathbb{R}^k\}$. This normal limiting-behaviour of estimator sequences in a neighbourhood of θ_0 may serve as an alternative explanation for the name 'local asymptotic normality'.

One caveat remains: the dependence on h of the limit distribution in (1.16) leaves room for highly irregular behaviour under small perturbations of the parameter or estimator sequence. As it turns out, this has especially grave consequences for efficiency and optimality. With regard to optimality the question is whether there are bounds to asymptotic estimation performance that are intrinsic to the estimation problem (*i.e.* which hold for *any* applied estimator sequence) and whether those bounds can actually be achieved. In the current situation, we have already assumed that T_n is consistent and converges at $\sqrt{n}$-rate. Further criteria for asymptotic performance are formulated in terms of the extent to which the limit distribution is concentrated around 0. Obviously there are many ways to quantify the 'degree

of concentration' (corresponding the risk under various choices for a loss-function), but we concentrate on asymptotic variance. Recall that an estimator sequence is said to be efficient if the variance of the limit distribution is minimal. Based on the arguments put forth thus far, this variance can be identified with the variance of a randomised statistic in the normal limit model in the case of a differentiable functional on a smooth, parametric model. More specifically, any estimator sequence T_n satisfying (1.16) estimates $\psi(\theta_0)$ consistently at rate $\sqrt{n}$ and $S_n = \sqrt{n}\big(T_n - \psi(\theta_0)\big)$ (as used in the proof of theorem 1.3) converges to S under $P_{\theta_0+h/\sqrt{n}}$. Hence an efficient estimator sequence T_n for $\psi(\theta_0)$ has a variance that equals the minimal variance of an estimator S for $\dot{\psi}_{\theta_0}h$ in the normal limit model. If we can determine the best estimator for $\dot{\psi}_{\theta_0}h$ in the normal model, we have a bound for optimal estimation of $\psi(\theta_0)$. Theorems concerning minimal-variance estimation in the normal model, however, involve additional requirements on the class of estimators under consideration. For instance, the Cramér-Rao theorem guarantees that the minimal variance equals the Fisher information *if* we restrict attention to the class of unbiased estimators. In the case at hand, suppose that $S = \dot{\psi}_{\theta_0}X$ (which is an unbiased estimator for $\dot{\psi}_{\theta_0}h$ since $X \sim N(h, I_{\theta_0}^{-1})$. Then $S - \dot{\psi}_{\theta_0}h \sim N(0, \dot{\psi}_{\theta_0}I_{\theta_0}^{-1}\dot{\psi}_{\theta_0}^T)$, which is of minimal variance within the class of unbiased estimators according to the Cramér-Rao bound. Taking $h = 0$, we see that the best possible performance of an 'asymptotically unbiased' estimator sequence T_n is characterised in terms of the minimal variance $\dot{\psi}_{\theta_0}I_{\theta_0}^{-1}\dot{\psi}_{\theta_0}^T$ for the limit law.

Another, more important restricted class in which optimality criteria can be formulated, is the class of so-called regular estimators, defined as follows: an estimator sequence T_n is said to be regular at θ_0 for estimation of $\psi(\theta_0)$ if the limit distribution appearing in (1.16) is independent of h, *i.e.* for all h:

$$\sqrt{n}\Big(T_n - \psi\big(\theta_0 + \frac{h}{\sqrt{n}}\big)\Big) \overset{\theta_0+h/\sqrt{n}}{\rightsquigarrow} L_{\theta_0}. \tag{1.19}$$

Regularity is also referred to as 'asymptotic equivariance-in-law' because the limit-distribution is invariant under 'shifts' in the model as parameterised by h. Applying this definition to the situation under consideration in theorem 1.3, we see that for a regular estimation sequence T_n,

$$S - \dot{\psi}_{\theta_0}h \overset{h}{\sim} L_{\theta_0},$$

for every $h \in \mathbb{R}^k$. For example, when we estimate θ_0 itself by an estimator sequence T_n such that $S = X$ in the normal model $\{N(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\}$, then $S - h \sim N(0, I_{\theta_0}^{-1})$ for every h, *i.e.* the limit distribution is independent of h, $L_{\theta_0,h} = L_{\theta_0}$, and T_n is regular.

Regular estimators have a number of very remarkable properties, most notably the so-called convolution theorem (see Hájek (1970) [43] and theorem 8.8 in Van der Vaart (1998) [91]), which characterises the limit distribution for regular estimator sequences as convolutions of the form $N(0, \dot{\psi}_{\theta_0}I_{\theta_0}^{-1}\dot{\psi}_{\theta_0}^T) * M_{\theta_0}$ for some probability distribution M_{θ_0}, which implies that the (co-)variance of L_{θ_0} is lower-bounded by $\dot{\psi}_{\theta_0}I_{\theta_0}^{-1}\dot{\psi}_{\theta_0}^T$. Also related to regularity is the so-called locally asymptotic minimax theorem (see Hájek (1972) [44]) which formulates a

lower-bound for the uniform risk in small neighbourhoods of the true parameter θ_0 (for a more precise formulation, see the original references or theorem 8.11 in Van der Vaart (1998) [91], pp. 117–118). In a one-dimensional model, if an estimator sequence satisfies the local asymptotic minimax criterion for a loss-function in a fairly general class, then this estimator sequence is also a best regular sequence (see lemma 8.13 in Van der Vaart (1998) [91]). Within the limited extent of this introductory overview, a more detailed discussion of the convolution and locally asymptotic minimax theorems would take us to far afield. We do mention an equivalent characterisation of best regular estimators, however, in the form of the following theorem.

Theorem 1.4. *(Best regular estimation) Assume that the model $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean at θ_0 with non-singular Fisher information matrix I_{θ_0}. Let T_n be statistics in the models $\{P_{\theta_0+h/\sqrt{n}} : h \in \mathbb{R}^k\}$ estimating a functional $\psi : \Theta \to \mathbb{R}$ (differentiable at θ_0). Then the following two assertions are equivalent:*

(i) The sequence T_n is best regular for estimation of $\psi(\theta_0)$.

(ii) The sequence T_n satisfies:

$$\sqrt{n}(T_n - \psi(\theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\psi}_{\theta_0} I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1).$$

Proof The proof of this theorem can be found in Van der Vaart (1998) [91], pp. 120–121. □

Comparing the above with expansion (1.13), we see that under the regularity conditions of theorem 1.1, the maximum-likelihood estimator is best regular. Already in the 1940's (see Cramér (1946) [22]) it was known that maximum-likelihood estimation is asymptotically optimal under differentiability and Lipschitz conditions that are now known as 'classical conditions'.

That a condition like regularity to restrict the class of estimators over which an asymptotic optimality criterion is formulated, is *necessary*, is established by the construction of explicit counterexamples. It is possible to construct estimator sequences that elude the lower bound for the asymptotic variance (or even display a higher rate of convergence) if the true parameter θ_0 happens to be a specific point in the model, while maintaining the optimal rate and asymptotic variance for all other points. This phenomenon is called super-efficiency for obvious reasons. The first super-efficient estimator was found by Hodges in 1951 and others have been constructed since, *e.g.* shrinkage estimators (see, for instance, James and Stein (1961) [49]). The above theorems show that super-efficient estimator sequences for differentiable functionals in smooth parametric models cannot be regular. Furthermore, it was shown by Le Cam (1953) [63] that super-efficiency in can occur only on a set of Lebesgue measure equal to zero (Le Cam's (first) proof relies on the Bernstein-Von-Mises theorem, which is discussed in section 1.3 and of chapter 2). Indeed a version of the convolution theorem exists that does not require regularity of the estimator sequence, but holds only Lebesgue almost-everywhere

on the parameter space (see Van der Vaart (1998) [91], theorem 8.9 and lemma 8.10). The conclusion has to be that on the one hand it is possible to improve on estimator performance on a selected set of points, but on the other hand, that this set cannot have Lebesgue-measure greater than zero and that the very basic requirement of regularity is enough to exclude this. The interested reader is referred to reviews on super-efficiency, *e.g.* Pollard (2001) [76] and Van der Vaart (1997) [90].

1.3 Bayesian asymptotics

Obviously the limit of infinite sample-size can be studied in the Bayesian context as well. In Bayesian procedures, we have on the one hand a prior Π on a model $\mathscr{P}$ and on the other the *i.i.d.* sample $X_1, X_2, \ldots$, which through conditioning defines a sequence of posterior probability measures $\Pi_n(\,\cdot \,|\, X_1, X_2, \ldots, X_n\,)$ on the model $\mathscr{P}$. The asymptotic behaviour of the Bayesian procedure concerns the way in which posterior measures concentrate their mass around a point (or set of points) of convergence. Note that the estimation procedure is entirely fixed as soon as we choose prior and model, so all conditions for theorems are formulated in terms of Π and $\mathscr{P}$. We have already noted in section 1.1 that the posterior measure allows for a number of derived point estimators like the posterior mean and median. The asymptotic properties of such point estimators can be related to the asymptotic properties of the posterior sequence on which they are based. Therefore bounds on rate and efficiency of point estimators such as those mentioned in section 1.2, have implications for the Bayesian procedure as well.

One can choose from several techniques to approach questions in Bayesian asymptotics; the approach we choose is based on test functions. Test functions and the properties that we need are discussed in more detail in the next section.

1.3.1 Bayesian consistency theorems

Consistency, certainly incontestable as an asymptotic criterion from the frequentist point of view, is not free of controversy in Bayesian statistics. Specifically, the subjectivist Bayesian point of view does not attach value to any special point of convergence P_0 because no 'underlying' or 'true' distribution for the sample $X_1, X_2, \ldots$ is assumed within the subjectivist paradigm. The notion of 'merging' is perhaps closer to the subjectivist's philosophy: given two different priors Π_1 and Π_2 on a model Θ, merging is said to occur if the total-variation distance between the posterior predictive distributions goes to zero (see Blackwell and Dubins (1962) [19] and, for an overview, Ghosh and Ramamoorthi (2003) [41]).

Here we choose a different point of view, which is essentially of a frequentist nature and motivation: we assume that the sample is *i.i.d.* P_0 and we assume moreover that $P_0 \in \mathscr{P}$ (for now). The question is whether the Bayesian procedure converges to the point P_0 in a suitable way. Relations between merging and posterior consistency as defined below are discussed in Diaconis and Freedman (1986) [24].

We start by defining consistency, Bayesian style. Let $\mathscr{P}$ be a model with topology $\mathscr{T}$ and prior Π on the corresponding Borel σ-algebra. Assume that $X_1, X_2, \ldots$ is an infinite *i.i.d.* sample from an unknown distribution $P_0 \in \mathscr{P}$. We say that the sequence of posterior measures $\Pi(\,\cdot\,|\,X_1, X_2, \ldots, X_n)$ is consistent if for every open neighbourhood U of P_0

$$\Pi_n(\,U\,|\,X_1, X_2, \ldots, X_n\,) \overset{P_0-a.s.}{\longrightarrow} 1. \tag{1.20}$$

In the case of a metric model $\mathscr{P}$ (with metric d), which covers all cases we shall consider, consistency is equivalent to the condition that for every $\epsilon > 0$:

$$\Pi_n(\,d(P, P_0) \geq \epsilon\,|\,X_1, X_2, \ldots, X_n\,) \overset{P_0-a.s.}{\longrightarrow} 0, \tag{1.21}$$

since the above display is the complement of an open ball and every open neighbourhood of P_0 contains an open ball centred on P_0.

Lemma 1.1. *Assume that $\mathscr{P}$ has a countable basis at P_0. Then definition (1.20) implies that the sequence of posterior measures Π_n on $\mathscr{P}$ converges weakly to the measure degenerate at P_0,*

$$\Pi_n(\,\cdot\,|\,X_1, X_2, \ldots, X_n\,) \rightsquigarrow \delta_{P_0}, \qquad (P_0 - a.s.). \tag{1.22}$$

If $\mathscr{P}$ is a normal space the converse holds as well.

Proof Assume that (1.20) holds for every open neighbourhood of P_0. Let the sequence U_k, $k \geq 1$ denote the countable basis[15] at P_0. Define for every $k \geq 1$ the set Ω_k such that $P_0^\infty(\Omega_k) = 1$ and the limit in (1.20) with $U = U_k$ holds on Ω_k. Note that $\Omega' = \cap_{k\geq 1}\Omega_k$ satisfies $P_0^\infty(\Omega') = 1$ and for all $\omega \in \Omega'$ and all $k \geq 1$:

$$\Pi_n\big(\,U_k \;\big|\; X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\,\big) \to 1, \qquad (n \to \infty).$$

Fix $\omega \in \Omega'$, let the open neighbourhood U of P_0 be given. Then U contains U_l for certain $l \geq 1$ and hence:

$$\Pi_n\big(\,U \;\big|\; X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\,\big) \geq \Pi_n\big(\,U_l \;\big|\; X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\,\big) \to 1$$

as $n \to \infty$. So the countable basis at P_0 ensures that (1.20) does not only hold P_0-almost-surely for each U separately, but P_0-almost-surely for all U simultaneously.

Let $(X_1, \ldots, X_n) = (X_1(\omega), \ldots, X_n(\omega))$ for some $\omega \in \Omega'$. Let $f : \mathscr{P} \to \mathbb{R}$ be bounded and continuous. Let $\eta > 0$ be given. Choose $M > 0$ such that $|f| \leq M$ and (using the continuity of f at P_0) let U be a neighbourhood of P_0 such that $|f(P) - f(P_0)| \leq \eta$ for all $P \in U$.

[15] A topological space $(X, \mathscr{T})$ has a countable basis at a point x, if there exists a sequence U_n, $(n \geq 1)$, of open neighbourhoods of x, such that every open neighbourhood U of x contains a U_l, for some $l \geq 1$. For this and other topological definitions used in this subsection, see, for example, Munkres (2000) [72], sections 32, 33 and 34.

Consider the absolute difference between the expectations with respect to the posterior and with respect to δ_{P_0}:

$$\begin{aligned}
&\Big|\int_{\mathscr{P}} f(P)\, d\Pi_n(P|X_1,\ldots,X_n) - f(P_0)\Big| \\
&\qquad \leq \int_{\mathscr{P}\setminus U} |f(P)-f(P_0)|\, d\Pi_n(P|X_1,\ldots,X_n) + \int_U |f(P)-f(P_0)|\, d\Pi_n(P|X_1,\ldots,X_n) \\
&\qquad \leq 2M\, \Pi_n(\,\mathscr{P}\setminus U \,|\, X_1, X_2, \ldots, X_n\,) + \sup_{P\in U} |f(P)-f(P_0)|\, \Pi_n(\,U \,|\, X_1, X_2, \ldots, X_n\,) \\
&\qquad \leq \eta + o(1), \qquad (n\to\infty).
\end{aligned}$$

We conclude that $\Pi_n(\,\cdot\,|\,X_1, X_2, \ldots, X_n\,) \rightsquigarrow \delta_{P_0}$, P_0-almost-surely.

Conversely, assume that the sequence of posteriors converges weakly to the Dirac-measure at P_0, P_0-almost-surely. Let U, an open neighbourhood of P_0, be given. Assuming that $\mathscr{P}$ is a normal space (which implies that $\mathscr{P}$ is T_0 and hence the singleton $\{P_0\}$ is closed), Urysohn's lemma (see Munkres (2000) [72], theorem 33.1) guarantees the existence of a continuous $f : \mathscr{P} \to [0,1]$ that separates the set $\{P_0\}$ from the (closed) complement of U, *i.e.* $f = 1$ at $\{P_0\}$ and $f = 0$ on $\mathscr{P}\setminus U$. Hence:

$$\begin{aligned}
\liminf_{n\to\infty} \Pi_n(\,U\,|\,X_1, X_2, \ldots, X_n\,) &= \liminf_{n\to\infty} \int_{\mathscr{P}} 1_U(P)\, d\Pi_n(P|X_1,\ldots,X_n) \\
&\geq \liminf_{n\to\infty} \int_{\mathscr{P}} f(P)\, d\Pi_n(P|X_1,\ldots,X_n) = \int_{\mathscr{P}} f(P)\, d\delta_{P_0}(P) = 1,
\end{aligned}$$

which holds P_0-almost-surely. □

A few remarks concerning the topological conditions are in order at this point. First of all, the essence of the property we use in the second part of the above proof (the existence of a continuous function separating points (or rather, (closed) singletons) from closed sets) is usually referred to as *complete regularity* (or the $T_{3\frac{1}{2}}$ property, between regularity (T_3) and normality (T_4)). Urysohn's lemma guarantees that normal spaces are completely regular, but strictly speaking the requirement of normality is too strong for the purpose. Secondly, metrisable spaces are first countable and normal, so the above implies the following corollary immediately.

Corollary 1.1. *If $\mathscr{P}$ is a model with metric d, (1.20), (1.21) and (1.22) are equivalent.*

The popular[16] condition that the model be a *separable* metric space (which is also sufficient to prove that weak convergence implies (1.20)) is therefore too strong. Note that compact Hausdorff spaces are normal and that regular spaces with a countable basis are metrisable by Urysohn's metrisation theorem.

Throughout the rest of this thesis, we shall be concerned with metric spaces $\mathscr{P}$ only, so we could also have chosen to specify lemma 1.1 to the metric situation immediately. The reason

[16] see, for instance, Ghosh and Ramamoorthi (2003) [41], p. 17.

for doing this in more detail is the introduction of the T_4 and $T_{3\frac{1}{2}}$ properties. In the following chapters, we shall make frequent use of so-called test-functions, whose conceptual purpose is that of Urysohn's separating function in a statistical setting, generalising the well-known hypothesis test based on critical regions. The analogy shall become clear in due course (for the less-than-patient, see condition (1.23)).

Point-estimators derived from a consistent Bayesian procedure are consistent themselves under some mild conditions. We reiterate that the notion of a point-estimator is not an entirely natural extension to the Bayesian framework: for example, if the model is non-convex, the expectation based on the posterior measure may lie outside the model. Similarly, perfectly well-defined posteriors may lead to ill-defined point-estimators due to integrability issues or non-existence of maximisers, which become more severe as the model becomes more complicated.

We endow the model $\mathscr{P}$ with the (restriction) of the norm-topology that follows from the total-variation norm $\|\,.\,\|$. We assume that the σ-algebra on $\mathscr{P}$ contains the corresponding Borel σ-algebra.

Theorem 1.5. *Assume that prior* Π *and underlying distribution* $P_0 \in \mathscr{P}$ *are such that the sequence of posteriors is consistent. Then the posterior mean* $\hat{P}_n$ *is a* P_0*-almost-surely consistent point-estimator with respect to total-variation.*

Proof Note that the domain of definition of the map $P \mapsto \|P - P_0\|$ extends to the convex hull $\mathrm{co}(\mathscr{P})$ of $\mathscr{P}$ (in the collection of all probability distributions on the sample space). Since $P \mapsto \|P - P_0\|$ is convex by virtue of the triangle inequality, Jensen's inequality (see, *e.g.* theorem 10.2.6 in Dudley (1989) [28]) says that the posterior mean $\hat{P}_n$ satisfies:

$$\|\hat{P}_n - P_0\| = \Big\| \int_{\mathscr{P}} P \, d\Pi_n(\, P \,|\, X_1, \ldots, X_n\,) - P_0 \Big\| \leq \int_{\mathscr{P}} \|P - P_0\| \, d\Pi_n(\, P \,|\, X_1, \ldots, X_n\,).$$

Since $P \rightsquigarrow P_0$ under the sequence of posterior laws $\Pi_n = \Pi_n(\,\cdot\,|\,X_1, \ldots, X_n\,)$ and the map $P \mapsto \|P - P_0\|$ is bounded and continuous in the norm-topology, we conclude that the *r.h.s.* in the above display converges to the expectation of $\|P - P_0\|$ under the limit law δ_{P_0}, which equals zero. Hence

$$\hat{P}_n \to P_0, \qquad (P_0 - a.s.).$$

in total variation. □

More generally, given an arbitrary convex metric d on the model $\mathscr{P}$, theorem 1.5 can be proved if the metric d is bounded on $\mathscr{P}$. Similar arguments can be used to demonstrate consistency for other classes of point estimators derived from a consistent sequence of posterior distributions, for example Le Cam's so-called *formal Bayes estimators* [67].

Having discussed the definition of Bayesian consistency and its consequences for derived point-estimators, we move on to sufficient conditions for consistency. Perhaps the most famous

consistency theorem in Bayesian statistics is that given by Doob as early as 1948 [27], which states the following.

Theorem 1.6. *(Doob's consistency theorem) Suppose that both the model Θ and the sample space $\mathscr{X}$ are Polish spaces endowed with their respective Borel-σ-algebras. Assume that the map $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ is one-to-one. Then the sequence of posteriors is consistent Π-almost-surely.*

Proof The proof of this theorem is an application of Doob's martingale convergence theorem and can be found in Van der Vaart (1998) [91] and in Ghosh and Ramamoorthi (2003) [41]. □

For many Bayesians, Doob's theorem is more than enough: for parametric models with a prior that dominates the restriction of the Lebesgue measure to Θ, the above theorem leaves room for inconsistency only on sets of Lebesgue measure zero. A popular way of stating this, is that consistency theorems like the above show that "the data overrides prior beliefs asymptotically". Conclusions like that should be drawn less readily and less strongly, however. First of all, consistency occurs only if the true distribution was not excluded from consideration in the first place by an ill-chosen prior. If the support of the prior does not contain the true distribution, inconsistency is guaranteed. In fact, this situation should be compared to that of ordinary model misspecification as discussed in section 1.4 and later chapters investigate exactly this situation.

But there is a more subtle point of criticism to be made: Doob's proof says nothing about specific points in the model, *i.e.* given a particular $P_0 \in \mathscr{P}$ underlying the sample, Doob's theorem does not give conditions that can be checked to see whether the Bayesian procedure will be consistent at this point in the model: it is always possible that P_0 belongs to the null-set for which inconsistency occurs. That, indeed, this may lead to grave problems, especially in non-parametric situations, becomes apparent when we consider some rather awkward (but nonetheless perfectly acceptable) counterexamples given by Freedman (1963,1965) [34, 35] and Diaconis and Freedman (1986) [24, 25]. Non-parametric examples of inconsistency in Bayesian regression can be found in Cox (1993) [21] and Diaconis and Freedman (1998) [26]. Basically what is shown is that the null-set on which inconsistency occurs in Doob's theorem can be rather large in non-parametric situations. Some authors are tempted to present the above as definitive proof of the fact that Bayesian statistics are useless in non-parametric estimation problems. More precise would be the statement that not every choice of prior is suitable and some may lead to unforeseen instances of inconsistency. The fact that they are unforeseen is related to the non-specific nature of the exceptional null-set in Doob's theorem. Fortunately, a theorem exists that provides sufficient conditions for consistency at a *specific* point $P_0 \in \mathscr{P}$.

Theorem 1.7. *(Schwartz' consistency theorem) Let $\mathscr{P}$ be a model with a metric d, dominated by a σ-finite measure μ and assume that this model is well specified: $P_0 \in \mathscr{P}$. Let Π be a prior on $\mathscr{P}$ and assume that the following two conditions hold:*

(i) For every $\eta > 0$,

$$\Pi\Big(P \in \mathscr{P} \,:\, -P_0 \log \frac{p}{p_0} \leq \eta \Big) > 0,$$

(ii) For every $\epsilon > 0$, there exists a sequence ϕ_n of test-functions such that:

$$P_0^n \phi_n \to 0, \qquad \sup_{\{P : d(P,P_0) > \epsilon\}} P(1 - \phi_n) \to 0. \tag{1.23}$$

Then all open neighbourhoods of P_0 have posterior measure equal to one asymptotically, i.e.:

$$\Pi_n\big(d(P, P_0) \geq \epsilon \,|\, X_1, X_2, \ldots, X_n \big) \overset{P_0-a.s.}{\longrightarrow} 0, \tag{1.24}$$

for all $\epsilon > 0$.

Proof Let $\epsilon > 0$ be given. Define V to be the complement of the open d-ball of radius ϵ around P_0 in $\mathscr{P}$:

$$V = \{\, P \in \mathscr{P} \,:\, d(P, P_0) \geq \epsilon \,\}.$$

We start by splitting the n-th posterior measure of V with the test function ϕ_n and taking the limes superior:

$$\begin{aligned} &\limsup_{n\to\infty} \Pi_n\big(V | X_1, \ldots, X_n\big) \\ &\qquad \leq \limsup_{n\to\infty} \Pi_n\big(V | X_1, \ldots, X_n\big)(1 - \phi_n) + \limsup_{n\to\infty} \Pi_n\big(V | X_1, \ldots, X_n\big)\phi_n. \end{aligned} \tag{1.25}$$

For given $\eta > 0$ (to be fixed at a later stage) we consider the subset $K_\eta = \{P \in \mathscr{P} : -P_0 \log(p/p_0) \leq \eta\}$. For every $P \in K_\eta$, the strong law of large numbers says that:

$$\Big| \mathbb{P}_n \log \frac{p}{p_0} - P_0 \log \frac{p}{p_0} \Big| \to 0, \quad (P_0 - a.s.).$$

Hence for every $\alpha > \eta$ and all $P \in K_\eta$, there exists an $N \geq 1$ such that for all $n \geq N$, $\prod_{i=1}^n (p/p_0)(X_i) \geq e^{-n\alpha}$, P_0^n-almost-surely. This can be used to lower-bound the denominator in the expression for the posterior P_0^n-almost-surely as follows:

$$\begin{aligned} \liminf_{n\to\infty} e^{n\alpha} \int_{\mathscr{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) &\geq \liminf_{n\to\infty} e^{n\alpha} \int_{K_\eta} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \\ &\geq \int_{K_\eta} \liminf_{n\to\infty} e^{n\alpha} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \geq \Pi(K_\eta), \end{aligned}$$

where we use Fatou's lemma to obtain the second inequality. Since by assumption, $\Pi(K_\eta) > 0$

we see that the first term on the *r.h.s.* of (1.25) can be estimated as follows:

$$\begin{aligned}
&\limsup_{n\to\infty} \Pi_n(V|X_1,\ldots,X_n)(1-\phi_n)(X_1,\ldots,X_n) \\
&\qquad = \limsup_{n\to\infty} \frac{\displaystyle\int_V \prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)(X_1,\ldots,X_n)\,d\Pi(P)}{\displaystyle\int_{\mathscr{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi(P)} \\
&\qquad \le \frac{\limsup_{n\to\infty} e^{n\alpha} \displaystyle\int_V \prod_{i=1}^n (p/p_0)(X_i)\,(1-\phi_n)(X_1,\ldots,X_n)\,d\Pi(P)}{\liminf_{n\to\infty} e^{n\alpha} \displaystyle\int_{\mathscr{P}} \prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(P)} \\
&\qquad \le \frac{1}{\Pi(K_\eta)} \limsup_{n\to\infty} f_n(X_1,\ldots,X_n),
\end{aligned} \tag{1.26}$$

where we use the following, P_0^∞-almost-surely defined sequence of non-negative random variables $(f_n)_{n\ge1}$, $f_n : \mathscr{X}^n \to \mathbb{R}$:

$$f_n(X_1,\ldots,X_n) = e^{n\alpha} \int_V \prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)(X_1,\ldots,X_n)\,d\Pi(P).$$

Fubini's theorem and the fact that the test-sequence can be assumed to be uniformly exponential (see lemma 1.2) guarantee that there exists a constant $\beta > 0$ such that for large enough n,

$$\begin{aligned}
P_0^\infty f_n = P_0^n f_n &= e^{n\alpha} \int_V P_0^n \Big(\prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)(X_1,\ldots,X_n) \Big)\, d\Pi(P) \\
&\le e^{n\alpha} \int_V P^n(1-\phi_n)\,d\Pi(P) \le e^{-n(\beta-\alpha)}.
\end{aligned} \tag{1.27}$$

We choose η strictly below β and can then choose α such that $\eta < \alpha < \frac{1}{2}(\beta+\eta)$. Markov's inequality can be used to show that:

$$P_0^\infty\big(f_n > e^{-\frac{n}{2}(\beta-\eta)}\big) \le e^{n(\alpha-\frac{1}{2}(\beta+\eta))}.$$

Hence the series $\sum_{n=1}^\infty P_0^\infty(f_n > \exp -\frac{n}{2}(\beta-\eta))$ converges and the first Borel-Cantelli lemma then leads to the conclusion that:

$$0 = P_0^\infty\Big(\bigcap_{N=1}^\infty \bigcup_{n\ge N} \big\{ f_n > e^{-\frac{n}{2}(\beta-\eta)} \big\} \Big) \ge P_0^\infty\Big(\limsup_{n\to\infty} \big(f_n - e^{-\frac{n}{2}(\beta-\eta)} \big) > 0 \Big)$$

Since $f_n \ge 0$, we see that $f_n \to 0$, $(P_0 - a.s.)$, which we substitute in (1.26).

We estimate the last term on the *r.h.s.* of (1.25) with an argument similar to that used above for the functions f_n. Note that $P_0^n \Pi(V|X_1,\ldots,X_n)\phi_n \le P_0^n\phi_n \le e^{-nC}$ for some

positive constant C, according to lemma 1.2. Markov's inequality and the first Borel-Cantelli lemma suffice to show that:

$$\phi_n \Pi\big(V | X_1, \ldots, X_n\big) \overset{P_0-a.s.}{\longrightarrow} 0. \tag{1.28}$$

Combination of (1.26) and (1.28) proves that (1.25) equals zero. Positivity of the posterior measure completes the proof of (1.24). □

Here we come back to the remarks on normality following the proof of lemma 1.1. Comparing condition (1.23) with the assertion of Urysohn's lemma or the definition of complete regularity, one notices conceptually similar roles for separating functions and test functions: the sequence of test functions in (1.23) 'separates' the singleton $\{P_0\}$ from the alternative, albeit as a stochastic, uniform limit.

The condition of domination in the above theorem is strictly speaking redundant: it is possible to give the entire proof in its present form, if we replace p/p_0 by the Radon-Nikodym derivative dP/dP_0 (see footnote 2) throughout and we change the third equality in (1.27) into less-or-equal. Furthermore, it should be noted that the examples of non-parametric Bayesian inconsistency given by Diaconis and Freedman mentioned earlier in this subsection fail the prior-mass condition for Kullback-Leibler neighbourhoods of the true distribution in Schwartz' theorem (see Barron *et al.* (1999) [5]). Questions concerning the conditions under which suitable sequences of tests exist are answered, for instance, in Birge (1983,1984) [15, 16], Le Cam (1986) [67], Van der Vaart (1998) [91] and Ghosal *et al.* (2000) [39]. We conclude this subsection with the lemma referred to earlier, which is a special case of lemma 2.6.

Lemma 1.2. *Suppose that for given $\epsilon > 0$ there exists a sequence of tests $(\phi_n)_{n\geq 1}$ such that:*

$$P_0^n \phi_n \to 0, \qquad \sup_{P \in V_\epsilon} P^n(1 - \phi_n) \to 0$$

where $V_\epsilon = \{P \in \mathscr{P} : d(P, P_0) \geq \epsilon\}$. Then there exists a sequence of tests $(\omega_n)_{n\geq 1}$ and positive constants C, D such that:

$$P_0^n \omega_n \leq e^{-nC}, \qquad \sup_{P \in V_\epsilon} P^n(1 - \omega_n) \leq e^{-nD} \tag{1.29}$$

1.3.2 Bayesian rates of convergence

Recalling the formulation of posterior consistency given in (1.21), we define the rate of convergence for a consistent sequence of posteriors as the maximal speed with which we can let the balls $d(P, P_0) < \epsilon$ shrink to radius zero, while still capturing a posterior mass that converges to one in the limit $n \to \infty$. We formalise this as follows. Again, let $\mathscr{P}$ be a model with metric d and prior Π. Assume that $X_1, X_2, \ldots$ is an infinite *i.i.d.* sample from an unknown distribution $P_0 \in \mathscr{P}$. Let the sequence ϵ_n be such that $\epsilon_n > 0$ and $\epsilon_n \downarrow 0$. We say that the sequence of posterior measures $\Pi(\,\cdot \,|\, X_1, X_2, \ldots, X_n)$ converges to P_0 (at least) at rate ϵ_n if for all sequences $M_n \to \infty$:

$$\Pi_n\big(\, d(P, P_0) \geq M_n \epsilon_n \,|\, X_1, X_2, \ldots, X_n \,\big) \xrightarrow{P_0} 0, \tag{1.30}$$

This definition differs from (1.21) in two respects: firstly, the radius of the balls is now n-dependent and secondly, the mode of convergence with respect to the distribution of the sample is in P_0-probability.

To demonstrate how this definition relates to the rate of convergence for derived point-estimators like the posterior mean, we consider the following. We assume that the σ-algebra on the model contains the Borel σ-algebra corresponding to the metric topology generated by d. We also assume that the sequence of posteriors satisfies (1.30). With the sequence ϵ_n, we define moreover the point estimators $\tilde{P}_n$ as (near-)maximisers in the model of the maps:

$$P \mapsto \Pi_n(\, B(P, \epsilon_n) \,|\, X_1, \ldots, X_n \,),$$

where $B(P, \epsilon) \subset \mathscr{P}$ is the d-ball of radius ϵ around P in the model.

Lemma 1.3. *For every sequence $M_n \to \infty$, the estimator sequence $\tilde{P}_n$ satisfies*

$$P_0^n\big(\, d(\tilde{P}_n, P_0) \leq 2M_n\epsilon_n \,\big) \to 1 \tag{1.31}$$

As a result, ϵ_n^{-1} is a lower bound for the rate at which $\tilde{P}_n$ converges to P_0 with respect to d.

Proof Let $\tilde{P}_n$ like above be given. By definition of a near-maximiser:

$$\begin{aligned}\Pi_n(\, B(\tilde{P}_n, M_n\epsilon_n) \,|\, X_1, \ldots, X_n \,) &\geq \sup_{P \in \mathscr{P}} \Pi(\, B(P, M_n\epsilon_n) \,|\, X_1, \ldots, X_n \,) - o_{P_0}(1) \\ &\geq \Pi(\, B(P_0, M_n\epsilon_n) \,|\, X_1, \ldots, X_n \,) - o_{P_0}(1).\end{aligned}$$

Because the first term on the *r.h.s.* of the above display converges to one (according to (1.30)) and the second to zero in P_0-probability, the *l.h.s.* converges to one in P_0-probability. Since $B(\tilde{P}_n, M_n\epsilon_n) \cap B(P_0, M_n\epsilon_n) = \emptyset$ if $d(\tilde{P}_n, P_0) > 2M_n\epsilon_n$, the fact that the total posterior mass of the model does not exceed one guarantees that $d(\tilde{P}_n, P_0) \leq 2M_n\epsilon_n$ with P_0-probability growing to one as $n \to \infty$, demonstrating that ϵ_n^{-1} is a lower bound to the rate. □

A proof that does not differ in an essential way from the above can be given for the centre point of the d-ball of minimal radius containing posterior mass $p > 1/2$. For the posterior mean we can prove a similar result if we specify the convergence of the posterior measure of complements of balls a little further. Consider a model $\mathscr{P}$ with Hellinger metric H and the corresponding Borel σ-algebra. By the convexity of $P \mapsto H^2(P, P_0)$, the fact that this map can be extended to the convex hull of $\mathscr{P}$ and Jensen's inequality (see the proof of theorem 1.5):

$$\begin{aligned}H^2(\hat{P}_n, P_0) = H^2\Big(\int_{\mathscr{P}} P \, d\Pi_n(\, P \,|\, X_1, \ldots, X_n \,), P_0\Big) &\leq \int_{\mathscr{P}} H^2(P, P_0) \, d\Pi_n(\, P \,|\, X_1, \ldots, X_n \,) \\ &= \int_{\{H(P,P_0) > M_n\epsilon_n\}} H^2(P, P_0) \, d\Pi_n(\, P \,|\, X_1, \ldots, X_n \,) \\ &\qquad + \int_{\{H(P,P_0) \leq M_n\epsilon_n\}} H^2(P, P_0) \, d\Pi_n(\, P \,|\, X_1, \ldots, X_n \,) \\ &\leq 2\Pi_n(\, H(P, P_0) > M_n\epsilon_n \,|\, X_1, \ldots, X_n \,) \\ &\qquad + M_n^2\epsilon_n^2 \Pi_n(\, H(P, P_0) \leq M_n\epsilon_n \,|\, X_1, \ldots, X_n \,)\end{aligned}$$

If $\Pi_n(\, H^2(P,P_0) > a_n^2 \,|\, X_1,\ldots,X_n\,) = o_{P_0}(a_n^2)$ for all (deterministic) sequences $a_n \downarrow 0$, the above display implies that $H(\hat{P}_n, P_0)$ is bounded in probability by a multiple of $M_n\epsilon_n$ (*c.f.* (1.31)) for all sequences $M_n \to \infty$, leading to the same conclusion as that of lemma 1.3.

The possibility to construct point estimator sequences from posterior distributions converging at the same rate (*e.g.* $\tilde{P}_n$ above), implies that limitations on the rate of convergence (arising in particular in non-parametric estimation problems, see (1.39) below, for example) derived for point estimation, apply unabated to Bayesian rates. This argument applies to other asymptotic performance criteria as well.

With regard to sufficient conditions for the defining property (1.30) of the Bayesian rate of convergence, we note that the number of references on this subject is relatively small when compared to the literature concerning Bayesian consistency. We note first of all, Le Cam (1973) [66] and Ibragimov and Has'minskii (1981) [47], who prove that under regularity conditions, posteriors on parametric models achieve $\sqrt{n}$-rate of convergence. We do not discuss their results here, because the next subsection deals with the more detailed Bernstein-Von Mises theorem which implies $\sqrt{n}$-rate of convergence. Le Cam (1986) [67] considers rates of convergence of formal Bayes estimators. Two references dealing with Bayesian rates of convergence in non-parametric models are Ghosal, Ghosh and Van der Vaart (2000) [39] and Shen and Wasserman (2001) [83]. We follow the approach of the former, not only here but also in chapter 3.

Again, we assume a (non-parametric) model $\mathscr{P}$ with metric d and prior Π. To formulate the main theorem of this subsection, we define, for every $\epsilon > 0$, a particular variant of the Kullback-Leibler neighbourhood used in Schwartz' theorem (theorem 1.7).

$$B(\epsilon) = \Big\{ P \in \mathscr{P} : -P_0 \log \frac{p}{p_0} \leq \epsilon^2,\, P_0\Big(\log \frac{p}{p_0}\Big)^2 \leq \epsilon^2 \Big\}. \tag{1.32}$$

Theorem 1.8. *Suppose that for a sequence ϵ_n such that $\epsilon_n > 0$, $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$, the following two conditions hold:*

(i) There exists a constant $C > 0$ such that:

$$\Pi\big(B(\epsilon_n)\big) \geq e^{-nC\epsilon_n^2}. \tag{1.33}$$

(ii) There exists a sequence ϕ_n of test-functions ϕ_n and a constant $L > 0$ such that:

$$P_0^n\phi_n \to 0, \qquad \sup_{P:d(P,P_0)\geq\epsilon_n} P^n(1-\phi_n) \leq e^{-nL\epsilon_n^2}. \tag{1.34}$$

Then for a sufficiently large $M > 0$,

$$P_0^n\Pi_n(\, d(P,P_0) \geq M\epsilon_n \,|\, X_1,\ldots,X_n\,) \to 0. \tag{1.35}$$

Note that the assertion establishes convergence in P_0-expectation, which implies convergence in P_0-probability, *c.f.* (1.30).

Proof Define, for every $\eta > 0$, $A(\eta) = \{P \in \mathscr{P} : d(P, P_0) \geq \eta\}$. The expectation in (1.35) can be decomposed using the tests ϕ_n; for every $n \geq 1$ and every $M > 1$, we have:

$$\begin{aligned} &P_0^n \Pi_n\big(A(M\epsilon_n) \,\big|\, X_1, \ldots, X_n \big) \\ &\quad = P_0^n\, \phi_n(X)\, \Pi_n\big(A(M\epsilon_n) \,\big|\, X_1, \ldots, X_n \big) + P_0^n (1-\phi_n)(X)\, \Pi_n\big(A(M\epsilon_n) \,\big|\, X_1, \ldots, X_n \big). \end{aligned}$$

We estimate the terms on the right-hand side separately. Due to the first inequality in (1.34), the first term converges to zero. To estimate the second term, we substitute (1.4) to obtain:

$$\begin{aligned} &P_0^n \Pi_n\big(A(M\epsilon_n) \,\big|\, X_1, \ldots, X_n \big)(1-\phi_n)(X) \\ &\quad = P_0^n \Big[\int_{A(M\epsilon_n)} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P)\, (1-\phi_n)(X) \Big/ \int_{\mathscr{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \Big] \end{aligned} \tag{1.36}$$

in which the denominator can be lower-bounded by application of lemma 1.4, since by assumption (1.33), $\Pi(B(\epsilon_n)) > 0$. Let Ω_n be the subset in $\mathscr{X}^n$ for which the inequality between left- and right-hand sides in the following display holds:

$$\int_{\mathscr{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \geq \int_{B(\epsilon_n)} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \geq e^{-(1+K)n\epsilon_n^2} \Pi(B(\epsilon_n)), \tag{1.37}$$

as in (1.40), with $K > 0$ as yet unspecified. Decomposing the P_0^n-expectation in (1.36) into separate integrals over Ω_n and $\mathscr{X}^n \setminus \Omega_n$, we find:

$$\begin{aligned} &P_0^n \Pi_n\big(A(M\epsilon_n) \,\big|\, X_1, \ldots, X_n \big)(1-\phi_n) \\ &\qquad \leq P_0^n \Pi_n\big(A(M\epsilon_n) \,\big|\, X_1, \ldots, X_n \big)(1-\phi_n) 1_{\Omega_n} + P_0^n(\mathscr{X}^n \setminus \Omega_n). \end{aligned}$$

Note that $P_0^n(\mathscr{X}^n \setminus \Omega_n) = o(1)$ as $n \to \infty$ according to (1.40). The first term is estimated as follows:

$$\begin{aligned} &P_0^n \Pi_n\big(A(M\epsilon_n) \,\big|\, X_1, \ldots, X_n \big)(1-\phi_n)\, 1_{\Omega_n} \\ &\qquad \leq \frac{e^{(1+K)n\epsilon_n^2}}{\Pi(B(\epsilon_n))} P_0^n \Big((1-\phi_n) \int_{A(M\epsilon_n)} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \Big) \\ &\qquad \leq \frac{e^{(1+K)n\epsilon_n^2}}{\Pi(B(\epsilon_n))} \int_{A(M\epsilon_n)} P^n\big(1-\phi_n\big)\, d\Pi(P) \\ &\qquad \leq e^{(1+K)n\epsilon_n^2} \frac{\Pi(A(M\epsilon_n))}{\Pi(B(\epsilon_n))} \sup_{P \in A(M\epsilon_n)} P^n\big(1-\phi_n\big), \end{aligned} \tag{1.38}$$

where we have substituted (1.37) and used the positivity of the integrand, applied Fubini's theorem and bounded the integrand by its supremum over $A(M\epsilon_n)$. Application of the second inequality in (1.34) gives:

$$P_0^n \Pi_n\big(A(M\epsilon_n) \,\big|\, X_1, \ldots, X_n \big)(1-\phi_n) \leq e^{(1+K+C-M^2L)n\epsilon_n^2} + o(1).$$

Hence, for all $K > 0$ there exists a constant $M > 0$ such that the above expression converges to zero. This leads us to conclude that:

$$P_0^n \, \Pi_n\big(A(M\epsilon_n) \,\big|\, X_1, \ldots, X_n \big) \to 0, \qquad (n \to \infty).$$

for sufficiently large $M > 0$. □

The rate theorem given here is a variation on theorem 2.1 in Ghosal, Ghosh and Van der Vaart (2000) [39]; their version is more general in two respects: first of all, they allow for a *sequence* of priors Π_n, replacing Π in definition (1.4). Secondly, they restrict attention to a sequence of models $\mathscr{P}_n$ that grows in Π_n-measure to the full model $\mathscr{P}$ (sufficiently fast). More importantly, however, instead of the condition requiring the existence of suitable test functions, they impose the following alternative condition:

(ii a) The ϵ-packing numbers[17] $D(\epsilon, \mathscr{P}_n, d)$ for the models $\mathscr{P}_n$ satisfy:

$$D(\epsilon_n, \mathscr{P}_n, d) \leq e^{n\epsilon_n^2}. \tag{1.39}$$

Under certain, fairly general conditions, this entropy condition implies the existence of a suitable sequence of test functions, as is shown in section 7 of Ghosal *et al.* (2000). (See also Birgé (1983,1984) [15, 16] and Le Cam (1986) [67].) As such, the entropy condition is less general than the version given in theorem 1.8. However, for most models entropy numbers are well-known or can be calculated, whereas the existence of suitable test sequences is certainly more involved. Furthermore, if d is the Hellinger metric, suitable test sequences exist and conditions like (1.39) are often viewed as representing the optimal rate of convergence. Under certain conditions (for instance, if likelihood ratios are bounded away from zero and infinity), optimality is proved in Birgé (1983) [15] and Le Cam (1973,1986) [66, 67]. Referring to chapter 3, condition (3.7) is of a similar nature and extensive explanation concerning the relation between test sequences and entropy can be found in section 3.6 (note that the choice $P^* = P_0$ renders the discussion given there applicable to well-specified models).

The attentive reader will have noticed that the condition $n\epsilon_n^2 \to \infty$ precludes $\sqrt{n}$-rates of convergence. So in its present form the theorem is unable to establish rate-optimality of Bayesian methods in many estimation problems, including those involving smooth parametric models as discussed in section 1.2. Indeed, application of the theorem in such situations typically leads to unnecessary $(\log n)^\alpha$-factors. Similarly, the theorem would lead to unnecessary logarithmic factors when applied to finite-dimensional sieves. At first sight, one might suspect that this is related to the nature of the assertion (1.35), which ascribes to ϵ_n only the role of an lower bound for the rate, but the situation is more complicated. As it turns

[17] The packing number $D(\eta, \mathscr{X}, \rho)$ of a space $\mathscr{X}$ with metric ρ is defined as the maximal number of points in $\mathscr{X}$ such that the ρ-distance between all pairs is at least η. This number is related to the so-called covering number $N(\eta, \mathscr{X}, \rho)$ which is defined as the minimal number of ρ-balls of radius η needed to cover $\mathscr{X}$, by the following inequalities: $N(\eta, \mathscr{X}, \rho) \leq D(\eta, \mathscr{X}, \rho) \leq N(\eta/2, \mathscr{X}, \rho)$.

out, $\sqrt{n}$-rates require localised[18] versions of conditions (1.33) and (1.39) and a more subtle sequence of estimations to replace the inequalities (1.38). However, the essential ingredients of the proof are unchanged and a comprehensive presentation of the (mostly technical) details here would conflict with the introductory nature of this chapter. The interested reader is referred to Ghosal *et al.* [39], section 5 and theorems 2.4, 7.1. More specifically with regard to $\sqrt{n}$-rates of convergence in smooth parametric models, we refer to the next subsection and chapter 2 in this thesis.

To conclude this section we give the lemma needed in the proof of theorem 1.8 to lower-bound the denominator of the posterior in probability.

Lemma 1.4. *Let $\epsilon > 0$ be given and let $B(\epsilon)$ be defined as in (1.32). If $\Pi(B(\epsilon)) > 0$, then for every $K > 0$:*

$$P_0^n\Big(\int_{B(\epsilon)} \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(P) \leq e^{-n\epsilon^2(1+K)}\Pi(B(\epsilon))\Big) \leq \frac{1}{nK^2\epsilon^2}. \tag{1.40}$$

Proof The proof of this lemma can be found as lemma 8.1 in Ghosal *et al.* [39] and follows from the proof of lemma 3.17 if we choose $P^* = P_0$. □

1.3.3 The Bernstein-Von-Mises theorem

Having considered consistency and rate of convergence in the previous two subsections, we turn to the Bayesian analog of the limit distribution next. More particularly, we prove the so-called Bernstein-Von-Mises theorem, which states that, under regularity conditions comparable to those we saw in subsection 1.2.2, the posterior distribution for $\sqrt{n}(\theta - \theta_0)$ converges (in a suitable sense) to a normal distribution with location Δ_{n,θ_0} (*c.f.* (1.18)) and covariance equal to the inverse Fisher information $I_{\theta_0}^{-1}$. The first results concerning limiting normality of a posterior distribution date back to Laplace (1820) ([62]). Later, Bernstein (1917) [4] and Von Mises (1931) [71] proved results to a similar extent. Le Cam used the term 'Bernstein-Von-Mises theorem' in 1953 [63] and proved its assertion in greater generality in relation to his work concerning super-efficiency (see the remarks made at the end of subsection 1.2.2). Walker (1969) [95] and Dawid (1970) [23] gave extensions to these results and Bickel and Yahav (1969) [12] proved a limit theorem for posterior means. The theorem as given here can be found in Van der Vaart (1998) [91] which follows (and streamlines) the presentation given in [68].

We consider an infinite, *i.i.d.* P_0-distributed sample $X_1, X_2, \ldots$ and a parametric, well-specified model $\{P_\theta : \theta \in \Theta\}$, where the parameter set Θ is open in $\mathbb{R}^k$ and with $\theta_0 \in \Theta$ such that $P_0 = P_{\theta_0}$.

[18] For a definition of what is meant by localisation in this context, see the remarks preceding lemma 3.10 and the conditions of theorem 3.8.

Theorem 1.9. *(Bernstein-Von-Mises) Let the model be differentiable in quadratic mean at θ_0 with non-singular Fisher information matrix I_{θ_0}. Suppose that for every $\epsilon > 0$ there exists a sequence of tests such that:*

$$P_{\theta_0}^n \phi_n \to 0, \qquad \sup_{\|\theta-\theta_0\|>\epsilon} P_\theta^n (1-\phi_n) \to 0. \tag{1.41}$$

Furthermore, suppose that in a neighbourhood of θ_0, the prior is dominated by the Lebesgue measure and that the corresponding density is continuous and strictly positive in θ_0. Then the sequence of posterior measures converges as follows:

$$\left\| \Pi_{\sqrt{n}(\theta-\theta_0)|X_1,\ldots,X_n} - N(\Delta_{n,\theta_0}, I_{\theta_0}^{-1}) \right\| \xrightarrow{P_0} 0.$$

Proof The proof of this theorem can be found in Van der Vaart (1998) [91], pp. 141–144. Alternatively, the reader may refer to theorem 2.1, which forms the misspecified version of the above (under more stringent conditions appropriate to the misspecified situation) and substitute $P^* = P_0$. □

The conditions in the above theorem again comprise a uniform testing condition separating θ_0 from the complements of balls and a condition that lower-bounds the prior mass of neighbourhoods of the true distribution[19]. Note that the testing condition (1.41) is considerably weaker than that of theorem 1.8 (*c.f.* (1.34)), whereas the conclusion of the Bernstein-Von-Mises theorem is stronger. The difference lies in the fact that theorem 1.8 applies to non-parametric and parametric models alike, whereas the above Bernstein-Von-Mises theorem is formulated only for smooth parametric models. Suitably differentiable Euclidean models allow for the definition of uniformly exponential (score-)test sequences which make it possible to extend tests against fixed alternatives to shrinking neighbourhoods (see, *e.g.* the proof of theorem 2.3). With reference to condition (1.39), we also note that (localised) covering numbers for Euclidean spaces grow like $(1/\epsilon)^d$ (where d is the dimension) with decreasing ϵ (see, for instance, the proof of lemma 4.15).

The (proof of the) Bernstein-Von-Mises theorem depends crucially on local asymptotic normality of the model at θ_0, as required through differentiability in quadratic mean (see the first condition of theorem 1.2). A heuristic explanation of the role of this model-property in the proof can be given as follows. Suppose that the prior is as required in the Bernstein-Von-Mises theorem. Then the posterior for the local parameter $H = \sqrt{n}(\theta - \theta_0)$ has a Lebesgue-density given by:

$$\begin{aligned}
&\pi_n(\, h \,|\, X_1, X_2, \ldots, X_n \,) \\
&\qquad = \prod_{i=1}^n p_{\theta_0 + h/\sqrt{n}}(X_i)\, \pi(\theta_0 + h/\sqrt{n}) \;\Big/\; \int \prod_{i=1}^n p_{\theta_0 + h'/\sqrt{n}}(X_i)\, \pi(\theta_0 + h'/\sqrt{n})\, dh',
\end{aligned}$$

[19] A sufficient condition could be given in the form of an n-dependent lower bound on the prior mass of a sequence of balls around θ_0 shrinking at rate $\sqrt{n}$, similar to (1.33). The stated condition, however, is sufficient and in most situations, more convenient.

P_0^n-almost-surely. Continuity of π at θ_0 implies that (up to an n-dependent proportionality constant), $\pi(\theta_0 + h/\sqrt{n})$ converges to the constant $\pi(\theta_0)$, which is strictly positive by assumption. This makes it plausible that upon substitution of the likelihood expansion (1.15), the posterior density converges to:

$$\frac{\prod_{i=1}^n p_{\theta_0+h/\sqrt{n}}(X_i)}{\int \prod_{i=1}^n p_{\theta_0+h'/\sqrt{n}}(X_i)\, dh'} \approx \frac{e^{h^T \Delta_{n,\theta_0} - \frac{1}{2} h^T I_{\theta_0} h}}{\int e^{h'^T \Delta_{n,\theta_0} - \frac{1}{2} h'^T I_{\theta_0} h'}\, dh'} \to \frac{dN(h, I_{\theta_0}^{-1})(\Delta)}{\int dN(h', I_{\theta_0}^{-1})(\Delta)\, dh'}$$

(in a suitable sense with respect to P_0). Here Δ is an observation in the normal limit model $\{N(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\}$ (Recall from the proof of theorem 1.3 that under P_0, Δ_{n,θ_0} converges weakly to $\Delta \sim N(0, I_{\theta_0}^{-1})$). The *l.h.s.* of the last display equals $dN(\Delta, I_{\theta_0}^{-1})(h)$ and is the posterior based on a sample consisting only of Δ and the (improper) Lebesgue prior for the limit model.

Regarding the assertion of the Bernstein-Von-Mises theorem, we note that the centring sequence Δ_{n,θ_0} for the normal limit sequence may be chosen differently. According to theorem 1.4, any best-regular estimator sequence $\tilde{\theta}_n$ for θ_0 satisfies:

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) = \Delta_{n,\theta_0} + o_{P_0}(1). \tag{1.42}$$

Since the total-variational distance $\|N(\mu, \Sigma) - N(\nu, \Sigma)\|$ is bounded by a multiple of $\|\mu - \nu\|$, we find that the assertion of the Bernstein-Von-Mises theorem can also be formulated with $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ replacing Δ_{n,θ_0}. Using the invariance of total-variation under rescaling and shifts, this leads to the conclusion that:

$$\left\| \Pi_{\theta|X_1,\ldots,X_n} - N(\tilde{\theta}_n, n^{-1} I_{\theta_0}^{-1}) \right\| \xrightarrow{P_0} 0,$$

for any best-regular estimator-sequence $\tilde{\theta}_n$. In particular, recall that according to theorem 1.1 and the limit (1.13), the maximum-likelihood estimator is best-regular under smoothness conditions on the (log-)likelihood. This serves to motivate the often-heard statement that "Bayesian statistics coincides with the maximum-likelihood estimator asymptotically". With regard to point estimators derived from the posterior distribution, we note that the model $\mathscr{P}$ can be regarded as a subset of the sphere of unity in the Banach space of all finite, signed measures on the sample space endowed with the total-variation norm. Any continuous functional f on this Banach space satisfies:

$$\left| f\big(\Pi_{\sqrt{n}(\,\theta - \theta^*)|X_1,\ldots,X_n}\big) - f\big(N(\Delta_{n,\theta_0}, I_{\theta_0}^{-1})\big) \right| \leq \|f\| \left\| \Pi_{\sqrt{n}(\,\theta - \theta^*)|X_1,\ldots,X_n} - N(\Delta_{n,\theta_0}, I_{\theta_0}^{-1}) \right\|,$$

and the *r.h.s.* of the above display converges to zero in P_0-probability by the Bernstein-Von-Mises theorem. So all point estimators that are also continuous functionals have asymptotic behaviour that is controlled by the Bernstein-Von Mises theorem.

The range of practical statistical problems for which a Bernstein-Von-Mises theorem or an argument of similar content is useful, extends far beyond the strict range of applicability

of theorem 1.9 above. More particularly, a general non- or semi-parametric version of the Bernstein-Von-Mises theorem does not exist (see, however, Shen (2002) [84]). Although it is sometimes claimed that Bernstein-Von-Mises-like results cannot hold in non-parametric models (see *e.g.*, Cox (1993) [21] and Freedman (1999) [36]), examples of locally normal limiting behaviour of the marginal posterior distribution for a marginal, finite-dimensional parameter of interest can be shown to occur in specific situations (see, for instance, recent work by Y. Kim and J. Lee [55, 56] on survival models). A more general theorem would be highly desirable and serve a wide range of applications: consider, for example, asymptotic confidence regions for efficient estimators in semi-parametric models (see, for example, Bickel, Ritov, Klaassen and Wellner [14]). According to a yet-to-be-devised semi-parametric Bernstein-Von-Mises theorem, confidence regions for efficient regular estimators (*e.g.* the maximum-likelihood estimator in many situations) asymptotically coincide with sets of high posterior probability. MCMC simulation of Bayesian procedures may thus lead to (numerical approximations to) asymptotic confidence regions for semi-parametric maximum-likelihood estimators.

1.4 Misspecification

Let us briefly revisit the first few steps in statistical estimation: as has been explained in the first section of this chapter, most statistical procedures start with the definition of a model $\mathscr{P}$ and immediately proceed to assume that this model choice realises the requirement of well-specification, *i.e.* it contains the distribution P_0 underlying the sample:

$$P_0 \in \mathscr{P}.$$

If the above display does not hold, the model is said to be misspecified. Since the sample (and with it, the distribution P_0) is given and unknown in statistical problems, the above display is a condition for the model rather than for P_0. However, in many cases it is *used* the other way around: a choice $\mathscr{P}$ for the model that imposes convenient properties for its elements (from mathematical, computational or practical point of view) together with well-specification, implies that P_0 itself displays those properties.

Usually a model choice is motivated by interpretability of the parameters in the description. For example, imagine a problem in which data is to be classified into N classes through estimation of the density: modelling each of the classes by a component of a discrete mixture of Gaussians on the sample space may lead to misspecification. Note that the number of interpretable classes is fixed and unimodality of the components implies that for each class there is a representative centre point. Similarly, the popularity of very simple parametric models (like normal families) lies not only in the simplicity of calculations within such a model, but also in the fact that it is parameterised by a location and a variance, both readily interpreted.

Whether out of convenience or interpretability, well-specified models are hard to motivate realistically. The only model that is guaranteed to be well specified is the fully non-parametric

model, *i.e.* the space of all probability distributions on the sample space. As soon as we impose any restrictions, we introduce a bias as well, and the larger the restriction (*i.e.* the smaller the model), the larger the bias. Whether or not such bias is problematic depends on the specific problem: if we are interested exclusively in some (rough) location estimate, a family of normal distributions may be misspecified yet suitable. If, on the other hand, we need a precise estimate of a certain density for predictive purposes, the same misspecified family of normals is wholly inadequate.

Many theorems in statistics involve well-specified models and surprisingly often, this assumption is far too strong. A more refined point of view arises when we dissociate the definition of the model from sufficient assumptions on P_0. Misspecified theorems are hence longer but state in far greater detail their maximal domain of applicability as far as the distribution of the data is concerned: given the model $\mathscr{P}$ and an *i.i.d.* P_0-distributed sample $X_1, \ldots, X_n$, the assertion holds if P_0 satisfies stated requirements. Usually, 'stated requirements' are satisfied by all $P \in \mathscr{P}$ and *in optima forma* also by a large set of other distributions, so that the misspecified theorem generalises its well specified version. In the present context, our primary interest is in theorems concerning consistency, rate and limit distribution of point estimation and Bayesian procedures.

1.4.1 Misspecification and maximum-likelihood estimation

One class of point estimators that generalises easily to the misspecified situation is that of M-estimators. Consider the sequence $\hat{\theta}_n$ of (near-)maximisers of the functions M_n over a model Θ, defined as in subsection 1.2.1. The argumentation around (1.9) holds in the misspecified case as well with one exception: the point θ^* at which $M(\theta)$ is maximal is such that $P_{\theta^*} \neq P_0$. If a unique maximum does not exist, the model is flawed more seriously than just by misspecification: in that case it is not P_0-identifiable (see section 3.3 for a formal definition). Roughly speaking, the existence of more than one maximum means that it is impossible to distinguish between them on the basis of a sample from P_0. Such situations do arise, for example if the model contains distinct P_1, P_2 differing only on a null-set of P_0. Identifiability issues are not specific to the misspecified situation: the fact that the function M must have a unique, well-separated maximum holds in the well-specified situation as well. Otherwise, the possibility arises that the sequence of maximum-likelihood estimators (if at all well-defined) does not converge to a point, but to a set of points.

For theorems concerning the asymptotic behaviour of estimation procedures, we therefore generalise θ^* like above to a point $P^* \in \mathscr{P}$ serving as an alternative point of convergence for the estimator sequence in the misspecified model. This alternative is the optimal approximation for P_0 within $\mathscr{P}$, in a sense that is to be specified further. As such, P^* may be loosely interpreted as a 'projection' of P_0 on the model $\mathscr{P}$, with uniqueness of the projection equivalent to identifiability under P_0.

Practically, the usefulness of theorems concerning misspecification should be clear: conclu-

sions regarding properties of the distribution that underlies the sample can only be trusted to the extent one is willing to trust the assumption of well-specification. Said conclusions are often drawn while consciously neglecting the fact that well-specification is rather hard to justify. Justification is nevertheless found in practical experience: in many cases the conclusion holds all the same, despite model misspecification. Apparently, the condition of well-specification was too strong, but exactly when this is the case remains unclear unless we formulate the same theorem in a misspecified setting.

Theoretically, there is an important area of application for theorems concerning misspecified asymptotics in the field of adaptive estimation. Suppose that P_0 is known to be in a model $\mathscr{P}$ with metric d. If we want this assumption to be weak, $\mathscr{P}$ will often have to be a large, non-parametric model. With reference to formula (1.39) it is noted that the rate of convergence may suffer. Therefore, we choose a sequence of nested submodels $\mathscr{P}_n \subset \mathscr{P}_{n+1} \subset \mathscr{P}$ and we choose the estimator $\hat{P}_n$ in $\mathscr{P}_n$. We define P_n^* to be such that:

$$d(P_n^*, P_0) = \inf_{P \in \mathscr{P}_n} d(P, P_0),$$

and we impose that $d(P_n^*, P_0) \to 0$ as $n \to \infty$, so $\mathscr{P}$ equals the closure of the union $\cup_n \mathscr{P}_n$. A sequence of submodels like this is called a sieve for estimation in $\mathscr{P}$. The metric distance from the estimator to P_0 satisfies:

$$d(\hat{P}_n, P_0) \leq d(\hat{P}_n, P_n^*) + d(P_n^*, P_0). \tag{1.43}$$

Since $\hat{P}_n$ is chosen within $\mathscr{P}_n$, the first term equals the rate of convergence appropriate for estimation within $\mathscr{P}_n$ (which may be as fast as $\sqrt{n}$ if the submodel is chosen 'small enough'). On the other hand, the rate at which the second term in the above converges to zero is higher when the submodels are chosen 'large enough'. A suitable choice of submodels $\mathscr{P}_n$ balances the two terms on the right (ideally), in the sense that they both go to zero at the same rate, which, ideally, is optimal for estimation of P_0 in $\mathscr{P}$ and subject to bounds like (1.39). Furthermore, if $P_0 \in \mathscr{P}_k$ for some fixed $k \geq 1$, the second term on the *r.h.s.* in the above display equals zero, and one would like to find that $\hat{P}_n$ converges to $P_n^* = P_k^* = P_0$ at a rate appropriate within the k-th submodel. The sequence $d(\hat{P}_n, P_n^*)$ is determined by the rate at which estimation in the submodel $\mathscr{P}_n$ can be made, with one important difference from 'standard' rate calculations: the model is misspecified, since $P_0 \notin \mathscr{P}_n$.

As an example of an estimation problem under misspecification, we consider maximum-likelihood estimation in a smooth parametric model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ without the assumption that $P_0 \in \mathscr{P}$. One quickly finds this problem to be an M-estimation problem if we choose the maximisation function m as:

$$m_\theta(X) = \log p_\theta(X).$$

This leads to a sequence of estimators $\hat{\theta}_n$ that are (near-)maximisers of $\theta \mapsto \mathbb{P}_n \log p_\theta$. Under conditions (*e.g.* Wald's conditions) for consistent M-estimation, the sequence $\hat{\theta}_n$ converges to

the point $\theta^* \in \Theta$ that minimises the so-called Kullback-Leibler divergence of P_θ with respect to P_0:

$$\theta \mapsto -P_0 \log \frac{p_\theta}{p_0},$$

over the model Θ. That θ^* does not correspond to the true distribution P_0 is inconsequential: the maximum-likelihood procedure defines the 'best' approximation of P_0 within $\mathscr{P}$ to be the point of minimal Kullback-Leibler divergence. In order to assess rate and limit distribution, note that we can still use theorem 1.1 which takes the following specific form.

Theorem 1.10. *For each $\theta \in \Theta$, let $x \mapsto \log p_\theta(x)$ be a measurable function such that $\theta \mapsto \log p_\theta(X)$ is P_0-almost-surely differentiable at θ^* with derivative $\dot{\ell}_{\theta^*}(X)$. Furthermore, suppose that there exists a P_0-square-integrable random variable $\dot{m}$ such that for all θ_1, θ_2 in a neighbourhood of θ^*:*

$$\left| \log \frac{p_{\theta_1}}{p_{\theta_2}}(X) \right| \leq \dot{m}(X)\, \|\theta_1 - \theta_2\|, \quad (P_0 - a.s.).$$

Let the expectations $\theta \mapsto -P_0 \log p_\theta$ have a second-order Taylor expansion around θ^*:

$$-P_0 \log \frac{p_\theta}{p_{\theta^*}} = \tfrac{1}{2}(\theta - \theta^*)^T V_{\theta^*}(\theta - \theta^*) + o(\|\theta - \theta^*\|^2), \quad (\theta \to \theta^*).$$

with non-singular second-derivative matrix V_{θ^*}. Then any sequence of estimators $\hat{\theta}_n$ such that $\hat{\theta}_n \xrightarrow{P_0} \theta^*$ and $\mathbb{P}_n \log p_{\hat{\theta}_n} \geq \sup_\theta \mathbb{P}_n \log p_\theta - o_{P_0}(n^{-1})$ satisfies:

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_{\theta^*}^{-1} \dot{\ell}_{\theta^*}(X_i) + o_{P_0}(1).$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta^*)$ is asymptotically normal with mean zero and covariance matrix:

$$V_{\theta^*}^{-1}\, P_0\, [\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T]\, V_{\theta^*}^{-1}.$$

Only in the last assertion of the theorem do we see a significant departure from the well-specified situation: where previously the inverse Fisher-information and the expectation of the square of the score eliminated each other, this does not happen in the misspecified situation. Ultimately, this is a consequence of the fact that in general (refer to footnote 11 in this chapter for an explanation):

$$P_0\Big(\frac{p_\theta}{p_{\theta^*}}\Big) \neq 1,$$

unless the model is well specified and θ^* is such that $P_{\theta^*} = P_0$, a fact that will also play a prominent role in the chapters that follow.

To summarise the conclusions concerning maximum-likelihood estimation, we consider consistency, rate and limit distribution separately. Obviously, a misspecified model does not leave room for consistency in the truest sense: at best, the model is dense in another model that does contain P_0 and estimation may be such that $d(\hat{P}_n, P_0) \xrightarrow{P_0} 0$. Generically, however, this distance does not go to zero and it is even possible that it does not even converge if more than

one point P^* in the model is 'closest' to P_0. Assuming that a unique 'best' approximation P^* for P_0 exists (or under sufficient conditions to that effect, like in M-estimation) 'consistency' in misspecified models means that $d(\hat{P}_n, P^*) \xrightarrow{P_0} 0$. With regard to rate and limit distribution, we note that the rate often remains $1/\sqrt{n}$ and the limit distribution is still normal in theorem 1.10, but the asymptotic variance may change.

1.4.2 Misspecification in Bayesian statistics

What is meant by misspecification in Bayesian statistics is not immediately clear, because it is the prior (and not its support) that is native to the Bayesian framework. However, a choice for the prior with a support that does not contain P_0 is the Bayesian analog of a misspecified model. Equivalently, the model is misspecified in the Bayesian sense if there exists a neighbourhood of P_0 with prior mass zero (note that this is dependent on the choice of topology).

Before we continue under the assumption of Bayesian misspecification (valid throughout the following chapters unless specified otherwise), we briefly consider more sophisticated alternatives for the choice of a prior. Instead of making an (un)educated guess at the model it is also possible to choose the prior based on the sample, a method known as *empirical Bayes* (a variation on this theme, called *hierarchical Bayes*, assumes a second prior on the space of possible choices for the prior, conditioning on the data to obtain a posterior). For more on this subject, see, for instance, Berger (1985) [6] or Ripley (1996) [79]. From an asymptotic point of view, this opens the interesting possibility to consider adaptive Bayesian procedures. Bayesian model selection (for an overview, see Kass and Raftery (1995) [52]), which chooses from a collection of priors with different supports and model averaging (which mixes by means of a posterior in such a collection), particularly on a sieve within a non-parametric model (for instance, adaptive density estimation in smoothness classes (see *e.g.* Birgé and Massart (1997) [17])) seem to be viable.

This level of refinement is not the setting of this discussion, however: we assume throughout the rest of this thesis that the choice of prior has been made and has a support that does not necessarily include P_0 (a possibility that cannot be discounted when doing empirical Bayes either). To gain some insight, we end this chapter with a numerical exploration of Bayesian statistics with misspecified models. More specifically, we consider a normal location model:

$$\mathscr{P} = \{\, N(\theta, 1) \,:\, \theta \in [-1, 2] \,\},$$

and prior Π with polynomial Lebesgue-density of the form:

$$\pi(\theta) = \tfrac{1}{2} - \tfrac{4}{18}(\theta - \tfrac{1}{2})^2, \qquad (\theta \in [-1, 2]).$$

The sample is drawn *i.i.d.* from a uniform distribution on the interval $[0, 2]$. Figure 1.1 shows the posterior density at growing sample sizes. Under the premise of an extension of the Bernstein-Von-Mises theorem to misspecified models, we formulate our expectations and

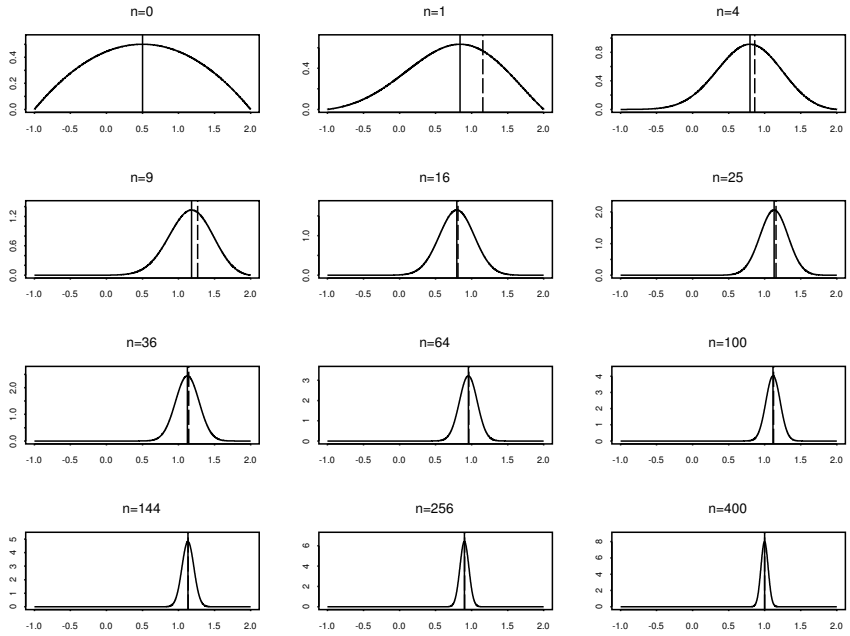

FIGURE 1.1 Convergence of the posterior density over a misspecified model. The samples, consisting of n observations, are *i.i.d.* uniform $U[0,2]$, the model consists of all normal distributions with mean between -1 and 2 and variance 1 and has a polynomial prior, shown in the first ($n = 0$) graph. The horizontal axis parameterises the location of the estimating normal distribution. On the vertical axis the posterior density is represented (note that the scale of the vertical axis varies). For all sample sizes, the *maximum a posteriori* and maximum likelihood estimators are indicated by a vertical line and a dashed vertical line respectively. Expectation and variance of the posterior distribution can be found in table 1.1.

discuss the extent to which they are realised here: if the model is suitably regular and the prior has a density that is continuous and strictly positive at θ^*, then we expect that posterior converges to a normal distribution like:

$$\left\| \Pi_{\theta|X_1,\ldots,X_n} - N(\tilde{\theta}_n, n^{-1}V_{\theta^*}) \right\| \xrightarrow{P_0} 0,$$

for a centring sequence $\tilde{\theta}_n$ and asymptotic covariance V_{θ^*}. The actual theorem forms the subject of chapter 2. First of all, we would expect consistency: the posterior should converge to a degenerate measure at θ^*, which appears to be happening for $\theta^* \approx 1.0$. From $n = 16$ onward, moreover, the posterior densities display a marked resemblance with normal densities of shrinking variance. Since convergence in total variation is equivalent to L_1-convergence of densities, this is to be expected from the Bernstein-Von-Mises theorem (although the small sample-sizes for which it obtains in this example are perhaps somewhat surprising). That, indeed, the rate of convergence $1/\sqrt{n}$ applies, becomes apparent when we evaluate the asymp-

totic variance of the posterior and rescale by a factor n (see table 1.1). Based on the close

n	$\Pi_n\theta$	$n\mathrm{Var}_{\Pi_n}(\theta)$
0	0.500	—
1	0.792	0.350
4	1.038	0.634
9	0.790	0.880
16	1.039	0.909
25	1.172	0.926

n	$\Pi_n\theta$	$n\mathrm{Var}_{\Pi_n}(\theta)$
36	0.995	0.964
64	0.835	0.984
100	0.993	0.988
144	0.964	0.992
256	0.997	0.995
400	1.005	0.997

TABLE 1.1 The table gives the sample size n, posterior expectation $\Pi_n\theta$ and n-rescaled posterior variance $n\Pi_n(\theta - \Pi_n\theta)^2$ corresponding to the samples that were used in figure 1.1.

relationship that exists between maximum-likelihood estimation and Bayesian methods and on the material presented at the end of the previous subsection, the point of convergence θ^* is expected to equal the location of the minimum of the Kullback-Leibler divergence. In this model,

$$-P_0 \log \frac{p_\theta}{p_0} = \tfrac{1}{2} P_0 (X - \theta)^2 + \text{constant}.$$

which is minimised by $\theta^* = P_0 X$. In the case at hand $P_0 = U[0, 2]$, so $\theta^* = 1$. This implies that:

$$-P_0 \log \frac{p_\theta}{p_{\theta^*}} = \tfrac{1}{2} V_{\theta^*} (\theta - \theta^*)^2$$

with $V_{\theta^*} = 1$. According to theorem 2.1, the variance of the normal distribution to which the posterior converges equals V_{θ^*}, in this case 1, in agreement with table 1.1.

With regard to the centring sequence we consider the relation between Bayesian and maximum-likelihood estimation. To that end, the MLE has been included in the graphs of figure 1.1 as well. Bridging the two concepts is the *maximum a-posteriori* estimator which is also represented in the graphs. Referring to formula (1.3) we see that the MAP estimator maximises[20]:

$$\theta \mapsto \prod_{i=1}^{n} p_\theta(X_i)\, \pi(\theta).$$

If the prior had been uniform, the last factor would have dropped out and the maximisation of the posterior density *is* maximisation of the likelihood, so differences between ML and MAP

[20] There is an interesting connection with penalised maximum likelihood estimation (see, Van de Geer (2000) [37]) here: Bayes' rule applied to the posterior density $\pi_n(\theta|X_1, \ldots, X_n)$ gives:

$$\log \pi_n(\theta|X_1, \ldots, X_n) = \log \pi_n(X_1, \ldots, X_n|\theta) + \log \pi(\theta) + \text{constant}.$$

The first term equals the log-likelihood and the logarithm of the prior plays the role of a penalty term when maximising over θ.

estimators are entirely due to non-uniformity of the prior. Asymptotically, non-uniformity of the prior becomes irrelevant and MAP and ML estimators converge, as indicated by the Bernstein-Von-Mises theorem: the likelihood product overwhelms the last factor in the above display as $n \to \infty$.

Chapter 2

The Bernstein-Von-Mises theorem under misspecification

The main result of this chapter is theorem 2.1, the analog of the Bernstein-Von-Mises theorem 1.9 under misspecification. The ordinary Bernstein-Von-Mises theorem 1.9 has three principle conditions: regularity of the model at θ_0, uniform testability of P_{θ_0} versus alternatives at fixed distance and sufficiency of prior mass in neighbourhoods of θ_0. We shall see that analogous conditions apply in the misspecified case. Regularity conditions to guarantee local asymptotic normality of the model under P_0 (see lemma 2.2) are slightly stronger than those found in theorem 1.9 and resemble the regularity conditions of theorems 1.1 and 1.10. The other principle condition of theorem 2.1 requires $\sqrt{n}$-rate of posterior convergence (see condition (2.6)). In later sections we show that this rate can be ensured by uniform testability and prior mass conditions analogous to those in theorem 1.9.

In view of the results given in chapter 3 on (non-parametric) posterior rates of convergence, it may seem strange that uniform testability of P_0 versus fixed alternatives is sufficient. As it turns out, the regularity properties formulated in lemma 2.2 *also* enable the extension of such tests to complements of shrinking balls. Locally, the construction relies on score-tests to separate the point of convergence from complements of neighbourhoods shrinking at rate $1/\sqrt{n}$, using Bernstein's inequality to obtain exponential power. Assumed tests for fixed alternatives are used to extend those local tests to the full model. Of course, the condition on the prior measure guarantees that the rate of convergence is not limited by sparsity of prior mass in neighbourhoods of the point of convergence (by comparison, see theorem 1.8 and, more specifically, condition (1.33)).

Finally, we give a fairly general lemma to exclude testable model subsets, which implies a misspecified version of Schwartz' consistency theorem as given in 1.7. The presentation of these results as found in the remainder of this chapter is to be submitted to the *Annals of Statistics* for publication.

The Bernstein-Von-Mises theorem under misspecification

B.J.K. KLEIJN AND A.W. VAN DER VAART

Free University Amsterdam

Abstract

We prove that the posterior distribution based on an *i.i.d.* sample in misspecified parametric models converges to a normal limit in total variation under conditions that are comparable to those in the well-specified situation. Besides regularity conditions, uniform testability against fixed alternatives and sufficiency of prior mass in neighbourhoods of the point of convergence are required. The rate of convergence is considered in detail, with special attention for the existence and construction of suitable test sequences. We also give a lemma to exclude testable model subsets which implies a misspecified version of Schwartz' consistency theorem, establishing weak convergence of the posterior to a measure degenerate at the point at minimal Kullback-Leibler divergence with respect to the true distribution.

2.1 Introduction

Basic estimation problems involving a sample of n observations $X_1, \ldots, X_n$ (assumed to be *i.i.d.* P_0), require an estimate $\hat{P}_n$ for P_0 in a model $\mathscr{P}$. Such problems involve certain assumptions about P_0 and require a definite choice for the model $\mathscr{P}$ unless some form of model-selection is used. Assumptions concerning P_0 often arise from the context of the experiment.Usually, the model choice is made on the basis of interpretability of the parameterisation. For instance, the ever-popular normal model leaves only the expectation and variance of the data to be estimated, both of which are readily interpreted. In less extreme cases, models are chosen as 'reasonable approximations' to P_0, where assumed properties of P_0 play a large role in the motivation for the approximation. Ultimately, $\hat{P}_n$ lies in $\mathscr{P}$, so in principle the choice of a model introduces a bias and, of course, the smaller the model, the larger this bias is.

These two aspects of statistical estimation, the choice for the model $\mathscr{P}$ and assumptions on the underlying distribution P_0 are often linked by the assumption that the model is *well specified*, *i.e.*

$$P_0 \in \mathscr{P}. \tag{2.1}$$

Properties of P_0 are then implied by the choice of $\mathscr{P}$ and the bias introduced by the choice of a model is assumed equal to zero. Although (2.1) can be very convenient from a mathematical point of view, otherwise the assumption is hard to justify. However, it is so common in mathematical statistics that it is omitted in the statement of theorems habitually. In applied statistics, theorems that rely on (2.1) are often used without mention of the fact that, in all likelihood, the true distribution of the data does *not* lie in the model.

The 'abuse' of well-specified theorems without proof of (2.1) in applications is motivated (and justified to a certain extent) by the fact that they often work regardless. This raises the question why, *i.e.* "Is it possible to prove those same theorems *without* the assumption $P_0 \in \mathscr{P}$?". This does not mean that no assumptions on P_0 are made, the point is merely to find restrictions on P_0 that delimit the theorem's range of applicability more appropriately. In this paper, this point of view is applied to the Bernstein-Von-Mises theorem and other theorems concerning the asymptotic behaviour of Bayesian procedures.

As is well-known, Bayesian procedures and maximum-likelihood estimation have a lot in common. Crude argumentation starts with the observation that both methods concentrate around regions in the model where the likelihood is high and hence one might expect that both minimise the Kullback-Leibler divergence asymptotically (see, for instance, Huber (1967) [46]). In well-specified models, this minimum obtains at the point P_0 in the model, so if the model is suitably identifiable and regular, one would expect that both methods lead to consistent estimates. Replacing P_0 in a well-specified model by a (unique) point P^* at which the Kullback-Leibler divergence with respect to P_0 is minimal, the heuristic relation between maximum-likelihood estimation and Bayesian procedures extends to the misspecified situation without fundamental differences.

The Bernstein-Von-Mises theorem says that in well-specified, smooth, parametric models, the analogy between ML estimation and Bayesian methods goes a lot further, asserting that a suitably centred and rescaled version of the posterior distribution converges in total variation to a normal limit distribution centred at the maximum-likelihood estimator with covariance equal to the inverse Fisher-information. This fact renders Bayesian credible sets and confidence regions for the maximum-likelihood estimator interchangeable asymptotically. The first results concerning limiting normality of a posterior distribution date back to Laplace (1820) [62]. Later, Bernstein (1917) [4] and Von Mises (1931) [71] proved results to a similar extent. Le Cam used the term 'Bernstein-Von-Mises theorem' in 1953 [63] and proved its assertion in greater generality. Walker (1969) [95] and Dawid (1970) [23] gave extensions to these results and Bickel and Yahav (1969) [12] proved a limit theorem for posterior means. A version of the theorem involving only first derivatives of the log-likelihood in combination with testability

and prior mass conditions (compare with Schwartz' consistency theorem, Schwartz (1965) [82]) can be found in Van der Vaart (1998) [91] which follows (and streamlines) the presentation given in [68]. More recently, posterior rates of convergence were considered in non-parametric models (see Ghosal, Ghosh and Van der Vaart (2000) [39] and Shen and Wasserman (2001) [83]), while semi-parametric versions of the Bernstein-Von-Mises theorem have been given in Shen (2002) [84] and by Kim and Lee [55, 56].

Based on the above heuristic argument, the sequence of posterior distributions in a misspecified model may be expected to concentrate its mass asymptotically in sets at minimal Kullback-Leibler divergence with respect to the true distribution P_0. Indeed results to this extent have been obtained (for early references on misspecification in Bayesian asymptotics, see Berk (1966,1970) [8, 9]). Posterior rates of convergence in misspecified non-parametric models were considered in Kleijn and Van der Vaart (2003) [57]. The misspecified version of the Bernstein-Von-Mises theorem is expected to involve the maximum-likelihood estimator as a centring sequence. Indeed, this conclusion is reached by Bunke and Milhaud (1998) [20]. Unfortunately, the conditions of theorem 4.1 therein are numerous and stringent and cannot be compared with conditions in the well-specified case.

Below we give a derivation of the Bernstein-Von-Mises theorem under misspecification that holds with conditions comparable to those given in Van der Vaart (1998) [91]. The presentation is split up along $\sqrt{n}$-rate of posterior convergence, which is a condition in the statement of the main theorem (theorem 2.1) and the assertion of theorem 2.2. Both rely on the same set of regularity conditions, as put forth in lemma 2.2, which establishes local asymptotic normality under misspecification. The extension of uniform tests for P_0 versus fixed alternatives to tests concerning growing alternatives is the subject of theorem 2.3. We conclude with a lemma (applicable in parametric and non-parametric situations alike) to exclude testable model subsets, which implies a misspecified version of Schwartz' consistency theorem.

2.2 Posterior limit distribution

2.2.1 Preliminaries

Let Θ be an open subset of $\mathbb{R}^d$ parameterising a model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$. For simplicity, we assume that there exists a single measure μ that dominates all P_θ and P_0, although our results seem to be generalisable to non-dominated parametric models. The density of P_0 with respect to μ is denoted p_0, the densities of P_θ are denoted p_θ. We assume that the Kullback-Leibler divergence with respect to P_0 is not infinite over the entire model and that it has a unique minimum $\theta^* \in \Theta$, *i.e.*:

$$-P_0 \log \frac{p_{\theta^*}}{p_0} = \inf_{\theta \in \Theta} -P_0 \log \frac{p_\theta}{p_0} < \infty. \tag{2.2}$$

The prior measure Π on Θ (with Borel sigma-algebra $\mathscr{A}$) is assumed to be a probability measure with Lebesgue-density π, continuous on a neighbourhood of θ^* and strictly positive at θ^*. Priors satisfying these criteria assign enough mass to (sufficiently small) balls around θ^* to allow for optimal (*i.e.* $\sqrt{n}$-) rates of convergence of the posterior if certain regularity conditions are met (see section 2.3). Although there are other, less stringent conditions, the above constitutes a sufficient and simple way to guarantee that the rate is not limited by the choice of prior, formulated in a fashion that can be regarded as 'natural' in a parametric context.

The posterior based on the n-th sample $(X_1, X_2, \ldots, X_n)$ is denoted $\Pi_n(\,\cdot\,|X_1, \ldots, X_n)$:

$$\Pi_n\big(\, A \;\big|\; X_1, \ldots, X_n \,\big) = \frac{\displaystyle\int_A \prod_{i=1}^{n} p_\theta(X_i)\pi(\theta)\, d\theta}{\displaystyle\int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\pi(\theta)\, d\theta}, \tag{2.3}$$

for $A \in \mathscr{A}$. To denote the random variable associated with the posterior distribution, we use the notation $\underline{\theta}$. In the majority of this paper, we 'localise' the model by centring on θ^* and rescaling by a factor of $\sqrt{n}$, introducing a parameter $H = \sqrt{n}(\underline{\theta} - \theta^*) \in \mathbb{R}^d$. The posterior for H (defined on a σ-algebra denoted $\mathscr{B}$), follows from that for θ by $\Pi_n(H \in B|X_1, \ldots, X_n) = \Pi_n(\sqrt{n}(\underline{\theta} - \theta^*) \in B|X_1, \ldots, X_n)$ for all $B \in \mathscr{B}$. Furthermore, for any measurable $K \subset \mathbb{R}^d$ such that $\Pi_n(H \in K|X_1, \ldots, X_n) > 0$, $\Pi_n^K(\,\cdot\,|X_1, \ldots, X_n)$ denotes the posterior conditional on K:

$$\Pi_n^K\big(\, B \;\big|\; X_1, \ldots, X_n \,\big) = \frac{\Pi_n(\, B \cap K \,|\, X_1, \ldots, X_n\,)}{\Pi_n(\, K \,|\, X_1, \ldots, X_n\,)}, \tag{2.4}$$

for all $B \in \mathscr{B}$.

2.2.2 Main result

The multivariate normal distribution located at $\Delta \in \mathbb{R}^d$ with covariance V is denoted $N_{\Delta,V}$. If Δ is a statistic or other random variable, the corresponding multivariate normal distribution is a random quantity as well. So given a fixed, invertible covariance matrix V_{θ^*}, the sequence

$$\Delta_{n,\theta^*} = V_{\theta^*}^{-1}\mathbb{G}_n\dot{\ell}_{\theta^*} \tag{2.5}$$

may serve as a (random) sequence of locations, where $\dot{\ell}_{\theta^*}$ denotes the score function at θ^* and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$ is the empirical process. The posterior distribution is a random measure as it depends on the sample through definition (2.3). Note that the assertion of the main theorem (theorem 2.1 below) involves convergence *in P_0-probability*, reflecting the sample-dependent nature of the two sequences of measures converging in total-variation norm.

Theorem 2.1. *Let the sample $X_1, X_2, \ldots$ be distributed i.i.d.-P_0. Let the model Θ, $\theta^* \in \Theta$ and prior Π be as indicated above. Assume that the Kullback-Leibler divergence and the log-likelihood satisfy the conditions of lemma 2.2 with invertible V_{θ^*}. Furthermore, assume that*

for every sequence of balls $(K_n)_{n\geq 1} \subset \mathbb{R}^d$ with radii $M_n \to \infty$, we have:

$$\Pi_n\big(\, H \in K_n \;\big|\; X_1, \ldots, X_n \,\big) \xrightarrow{P_0} 1. \tag{2.6}$$

Then the sequence of posteriors converges to a sequence of normal distributions in total variation:

$$\sup_{B\in\mathscr{B}} \Big| \, \Pi_n\big(\, H \in B \;\big|\; X_1, \ldots, X_n \,\big) - N_{\Delta_{n,\theta^*}, V_{\theta^*}}(B) \, \Big| \xrightarrow{P_0} 0. \tag{2.7}$$

Proof The proof is split into two parts: in the first part, we prove the assertion conditional on an arbitrary compact set $K \subset \mathbb{R}^d$ and in the second part we use this to prove (2.7). Throughout the proof we denote the posterior for H given $X_1, X_2, \ldots$ by Π_n and the normal distribution $N_{\Delta_{n,\theta^*}, V_{\theta^*}}$ by Φ_n. For $K \in \mathbb{R}^d$, conditional versions (*c.f.* (2.4)) are denoted Π_n^K and Φ_n^K respectively.

Let $K \subset \mathbb{R}^d$ be a compact subset of $\mathbb{R}^d$. For every open neighbourhood $U \subset \Theta$ of θ^* there exists an $N \geq 1$ such that for all $n \geq N$, $\theta^* + K/\sqrt{n} \subset U$. Since θ^* is an internal point of Θ, we can define, for large enough n, the random functions $f_n : K \times K \to \mathbb{R}$ by:

$$f_n(g,h) = \Big(1 - \frac{\phi_n(h)}{\phi_n(g)} \frac{s_n(g)}{s_n(h)} \frac{\pi_n(g)}{\pi_n(h)}\Big)_+,$$

where $\phi_n : K \to \mathbb{R}$ is the Lebesgue density of the (randomly located) distribution $N_{\Delta_{n,\theta^*}, V_{\theta^*}}$ (with V_{θ^*} as in (2.13)), $\pi_n : K \to \mathbb{R}$ is the Lebesgue density of the prior for the centred and rescaled parameter H and $s_n : K \to \mathbb{R}$ equals the likelihood product:

$$s_n(h) = \prod_{i=1}^n \frac{p_{\theta^* + h/\sqrt{n}}}{p_{\theta^*}}(X_i).$$

Since the conditions of lemma 2.2 are met by assumption, we have for every random sequence $(h_n)_{n\geq 1} \subset K$:

$$\log s_n(h_n) = h_n \mathbb{G}_n \dot{\ell}_{\theta^*} - \tfrac{1}{2} h_n^T V_{\theta^*} h_n + o_{P_0}(1),$$

$$\log \phi_n(h_n) = -\tfrac{1}{2}(h_n - \Delta_{n,\theta^*})^T V_{\theta^*}(h_n - \Delta_{n,\theta^*}) + \text{constant}.$$

For any two sequences $(h_n)_{n\geq 1}$, $(g_n)_{n\geq 1} \subset K$, $\pi_n(g_n)/\pi_n(h_n) \to 1$ as $n \to \infty$. Combining this with the above display and (2.5), we see that:

$$\begin{aligned}
&\log \frac{\phi_n(h_n)}{\phi_n(g_n)} \frac{s_n(g_n)}{s_n(h_n)} \frac{\pi_n(g_n)}{\pi_n(h_n)} \\
&\qquad = -h_n \mathbb{G}_n \dot{\ell}_{\theta^*} + \tfrac{1}{2} h_n^T V_{\theta^*} h_n + g_n \mathbb{G}_n \dot{\ell}_{\theta^*} - \tfrac{1}{2} g_n^T V_{\theta^*} g_n + o_{P_0}(1) \\
&\qquad\qquad - \tfrac{1}{2}(h_n - \Delta_{n,\theta^*})^T V_{\theta^*}(h_n - \Delta_{n,\theta^*}) + \tfrac{1}{2}(g_n - \Delta_{n,\theta^*})^T V_{\theta^*}(g_n - \Delta_{n,\theta^*}) \\
&\qquad = o_{P_0}(1)
\end{aligned}$$

as $n \to \infty$. Since $x \mapsto (1 - e^x)_+$ is continuous on $\mathbb{R}$, we conclude that for every pair of random sequences $(g_n, h_n)_{n\geq 1} \subset K \times K$:

$$f_n(g_n, h_n) \xrightarrow{P_0} 0, \quad (n \to \infty).$$

For fixed, large enough n, P_0^n-almost-sure continuity of $(g,h) \mapsto \log s_n(g)/s_n(h)$ on $K \times K$ is guaranteed by the Lipschitz-condition (2.12), since it implies that for all $g_1, g_2, h_1, h_2 \in K$:

$$\begin{aligned}\Big|\log\frac{s_n(g_1)}{s_n(h_1)} - \log\frac{s_n(g_2)}{s_n(h_2)}\Big| &\leq \Big|\log\frac{s_n(g_1)}{s_n(g_2)}\Big| + \Big|\log\frac{s_n(h_1)}{s_n(h_2)}\Big| \\ &\leq \sqrt{n}\,\mathbb{P}_n m_{\theta^*}\left(\|g_1 - g_2\| + \|h_1 - h_2\|\right), \qquad (P_0^n - a.s.),\end{aligned}$$

and tightness of m_{θ^*} suffices. As a result of lemma 2.3, we see that:

$$\|\Delta_{n,\theta^*}\| \leq \|V_{\theta^*}^{-1}\|\,\|\mathbb{G}_n\dot{\ell}_{\theta^*}\| = \sqrt{n}\|V_{\theta^*}^{-1}\|\,\|(\mathbb{P}_n - P_0)\dot{\ell}_{\theta^*}\| \leq \sqrt{n}\|V_{\theta^*}^{-1}\|\,\mathbb{P}_n m_{\theta^*},$$

P_0^n-almost-surely. Hence the location of the normal distribution $N_{\Delta_{\theta^*},V_{\theta^*}}$ is P_0^n-tight and we see that $(g,h) \mapsto \phi_n(g)/\phi_n(h)$ is continuous on all of $K \times K$ P_0^n-almost-surely as well. Continuity (in a neighbourhood of θ^*) and positivity of the prior density guarantee that this holds for $(g,h) \mapsto \pi_n(g)/\pi_n(h)$ as well. We conclude that for large enough n, the random functions f_n are continuous on $K \times K$, P_0^n-almost-surely. Application of lemma 2.10 then leads to the conclusion that:

$$\sup_{g,h\in K} f_n(g,h) \xrightarrow{P_0} 0, \quad (n \to \infty). \tag{2.8}$$

Assume that K contains a neighbourhood of 0 (which guarantees that $\Phi_n(K) > 0$) and let Ξ_n denote the event that $\Pi_n(K) > 0$. Let $\eta > 0$ be given and based on that, define the events:

$$\Omega_n = \big\{\omega : \sup_{g,h\in K} f_n(g,h) \leq \eta\big\}.$$

Consider the expression (recall that the total-variation norm $\|\cdot\|$ is bounded by 2):

$$P_0^n\big\|\Pi_n^K - \Phi_n^K\big\|1_{\Xi_n} \leq P_0^n\big\|\Pi_n^K - \Phi_n^K\big\|1_{\Omega_n\cap\Xi_n} + 2P_0^n(\Xi_n \setminus \Omega_n). \tag{2.9}$$

As a result of (2.8) the latter term is $o(1)$ as $n \to \infty$. The remaining term on the *r.h.s.* can be calculated as follows:

$$\begin{aligned}\tfrac{1}{2}P_0^n\big\|\Pi_n^K - \Phi_n^K\big\|1_{\Omega_n\cap\Xi_n} &= P_0^n\int\Big(1 - \frac{d\Phi_n^K}{d\Pi_n^K}\Big)_+ d\Pi_n^K\,1_{\Omega_n\cap\Xi_n} \\ &= P_0^n\int_K\Big(1 - \phi_n^K(h)\frac{\int_K s_n(g)\pi_n(g)dg}{s_n(h)\pi_n(h)}\Big)_+ d\Pi_n^K(h)\,1_{\Omega_n\cap\Xi_n} \\ &= P_0^n\int_K\Big(1 - \int_K \frac{s_n(g)\pi_n(g)\phi_n^K(h)}{s_n(h)\pi_n(h)\phi_n^K(g)}d\Phi_n^K(g)\Big)_+ d\Pi_n^K(h)\,1_{\Omega_n\cap\Xi_n}.\end{aligned}$$

Note that for all $g,h \in K$, $\phi_n^K(h)/\phi_n^K(g) = \phi_n(h)/\phi_n(g)$, since on K ϕ_n^K differs from ϕ_n only by a normalisation factor. We use Jensen's inequality (with respect to the Φ_n^K-expectation) for the (convex) function $x \mapsto (1-x)_+$ to derive:

$$\begin{aligned}\tfrac{1}{2}P_0^n\big\|\Pi_n^K - \Phi_n^K\big\|1_{\Omega_n\cap\Xi_n} &\leq P_0^n\int\Big(1 - \frac{s_n(g)\pi_n(g)\phi_n(h)}{s_n(h)\pi_n(h)\phi_n(g)}\Big)_+ d\Phi_n^K(g)\,d\Pi_n^K(h)1_{\Omega_n\cap\Xi_n} \\ &\leq P_0^n\int \sup_{g,h\in K} f_n(g,h)1_{\Omega_n\cap\Xi_n} d\Phi_n^K(g)\,d\Pi_n^K(h) \leq \eta.\end{aligned}$$

Combination with (2.9) shows that for all compact $K \subset \mathbb{R}^d$ containing a neighbourhood of 0,

$$P_0^n \big\| \Pi_n^K - \Phi_n^K \big\| 1_{\Xi_n} \to 0.$$

Now let $(K_m)_{m\geq 1}$ be a sequence of balls centred at 0 with radii $M_m \to \infty$. For each $m \geq 1$, the above display holds, so if we choose a sequence of balls $(K_n)_{n\geq 1}$ that traverses the sequence K_m slowly enough, convergence to zero can still be guaranteed. Moreover, the corresponding events $\Xi_n = \{\omega : \Pi_n(K_n) > 0\}$ satisfy $P_0^n(\Xi_n) \to 1$ as a result of (2.6). We conclude that there exists a sequence of radii $(M_n)_{n\geq 1}$ such that $M_n \to \infty$ and

$$P_0^n \big\| \Pi_n^{K_n} - \Phi_n^{K_n} \big\| \to 0, \tag{2.10}$$

(where it is understood that the conditional probabilities on the *l.h.s.* are well-defined on sets of probability growing to one). Combining (2.6) and lemma 2.12, we then use lemma 2.11 to conclude that:

$$P_0^n \big\| \Pi_n - \Phi_n \big\| \to 0,$$

which implies (2.7). □

Regarding the centring sequence Δ_{n,θ^*} and its relation to the maximum-likelihood estimator, we note the following straightforward lemma concerning the limit distribution of maximum-likelihood sequences.

Lemma 2.1. *Assume that the model satisfies the conditions of lemma 2.2 with non-singular V_{θ^*}. Then a sequence of estimators $\hat{\theta}_n$ such that $\hat{\theta}_n \xrightarrow{P_0} \theta^*$ and*

$$\mathbb{P}_n \log p_{\hat{\theta}_n} \geq \sup_\theta \mathbb{P}_n \log p_\theta - o_{P_0}(n^{-1})$$

satisfies the asymptotic expansion:

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n V_{\theta^*}^{-1} \dot{\ell}_{\theta^*}(X_i) + o_{P_0}(1). \tag{2.11}$$

Proof The proof of this lemma is a more specific version of the proof found in Van der Vaart (1998) [91] on page 54. □

First of all, the above implies that for maximum-likelihood estimators that converge to θ^*, the sequence $\sqrt{n}(\hat{\theta}_n - \theta^*)$ has a normal limit distribution with mean zero and covariance matrix $V_{\theta^*}^{-1} P_0[\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T] V_{\theta^*}^{-1}$. More important for present purposes, however, is the fact that according to (2.11), this sequence differs from Δ_{n,θ^*} only by a term of order $o_{P_0}(1)$. Since the total-variational distance $\|N_{\mu,\Sigma} - N_{\nu,\Sigma}\|$ is bounded by a multiple of $\|\mu - \nu\|$ as $(\mu \to \nu)$, the assertion of the Bernstein-Von-Mises theorem can also be formulated with the sequence $\sqrt{n}(\hat{\theta}_n - \theta^*)$ as the locations for the normal limit sequence. Using the invariance of total-variation under rescaling and shifts, this leads to the conclusion that:

$$\sup_{A \in \mathscr{A}} \Big| \Pi_n\big(\underline{\theta} \in A \;\big|\; X_1, \ldots, X_n \big) - N_{\hat{\theta}_n, n^{-1} V_{\theta^*}}(A) \Big| \xrightarrow{P_0} 0,$$

which, perhaps, demonstrates the usual interpretation of the Bernstein-Von-Mises theorem more clearly: the sequence of posteriors resembles more and more closely a sequence of 'sharpening' normal distributions centred at the maximum-likelihood estimators. More generally, any sequence of estimators satisfying (2.11) may be used to centre the normal limit sequence on. Note that the conditions for lemma 2.1, which derive directly from a fairly general set of conditions for efficiency in parametric M-estimation (see, theorem 5.23 in Van der Vaart (1998) [91]), are very close to the conditions of the above Bernstein-Von-Mises theorem.

Condition (2.6) fixes the rate of convergence of the posterior distribution (see Kleijn and Van der Vaart (2003) [57] for general results on posterior rates in misspecified (parametric and non-parametric) models). Since the balls K_n correspond to balls in Θ that shrink to $\{\theta^*\}$ at a rate (slightly slower than) $1/\sqrt{n}$, condition (2.6) can be stated equivalently by requiring that the posterior converges at rate $1/\sqrt{n}$ to a Dirac measure degenerate at θ^*. Sufficient conditions are given in section 2.3.

2.2.3 Misspecification and local asymptotic normality

Asymptotic normality of the sequence of posterior distributions depends crucially on local asymptotic normality of the model. Lemmas that establish this property (for and overview, see, for instance Van der Vaart (1998) [91]) usually assume a well-specified model, whereas current interest requires local asymptotic normality in misspecified situations. To that end we consider the following lemma which gives sufficient conditions.

Lemma 2.2. *If the function $\theta \mapsto \log p_\theta(X)$ is differentiable at θ^* ($P_0 - a.s.$) (with derivative $\dot{\ell}_{\theta^*}$) and:*

(i) there is an open neighbourhood U of θ^* and a non-negative, square-integrable random variable m_{θ^*} such that for all $\theta_1, \theta_2 \in U$:

$$\left| \log \frac{p_{\theta_1}}{p_{\theta_2}} \right| \leq m_{\theta^*} \|\theta_1 - \theta_2\|, \quad (P_0 - a.s.), \tag{2.12}$$

(ii) the Kullback-Leibler divergence with respect to P_0 has a second-order Taylor-expansion around θ^*:

$$-P_0 \log \frac{p_\theta}{p_{\theta^*}} = \tfrac{1}{2}(\theta - \theta^*) V_{\theta^*} (\theta - \theta^*) + o(\|\theta - \theta^*\|^2), \quad (\theta \to \theta^*), \tag{2.13}$$

where V_{θ^*} is the positive-definite $d \times d$-matrix of second derivatives with respect to θ of $-P_0 \log(p_\theta / p_{\theta^*})$ evaluated in θ^*.

then, for every random sequence $(h_n)_{n \geq 1}$ in $\mathbb{R}^d$ that is bounded in P_0-probability:

$$\log \prod_{i=1}^{n} \frac{p_{\theta^* + h_n/\sqrt{n}}}{p_{\theta^*}}(X_i) = h_n^T \mathbb{G}_n \dot{\ell}_{\theta^*} - \tfrac{1}{2} h_n^T V_{\theta^*} h_n + o_{P_0}(1) \tag{2.14}$$

Proof Using lemma 19.31 in Van der Vaart (1998) [91] for $\ell_\theta(X) = \log p_\theta(X)$, the conditions of which are satisfied by assumption, we see that for any sequence $(h_n)_{n\geq 1}$ that is bounded in P_0-probability:

$$\mathbb{G}_n\Big(\sqrt{n}\big(\ell_{\theta^*+(h_n/\sqrt{n})} - \ell_{\theta^*}\big) - h_n^T\dot{\ell}_{\theta^*}\Big) \xrightarrow{P_0} 0. \tag{2.15}$$

Hence, we see that

$$n\mathbb{P}_n \log\frac{p_{\theta^*+h_n/\sqrt{n}}}{p_{\theta^*}} - \mathbb{G}_n h_n^T\dot{\ell}_{\theta^*} - nP_0\log\frac{p_{\theta^*+h_n/\sqrt{n}}}{p_{\theta^*}} = o_{P_0}(1),$$

Using the second-order Taylor-expansion (2.13):

$$P_0\log\frac{p_{\theta^*+h_n/\sqrt{n}}}{p_{\theta^*}} - \frac{1}{2n}h_n^T V_{\theta^*}h_n = o_{P_0}(1),$$

and substituting the log-likelihood product for the first term, we find (2.14). □

First of all, it should be noted that other formulations[1] may suffice to prove (2.14) as well. However, the regularity conditions in this lemma are also used in a number of other places throughout this paper, for example to prove the existence of suitable test sequences in subsection 2.3.2. Secondly, both differentiability of densities and the Lipschitz condition on log-likelihoods can be controlled by a suitable choice of the model and some mild assumptions concerning P_0. It seems that the most restrictive condition (in the sense that it limits the set of P_0 that are suited) is the existence of a second-order Taylor-expansion for the Kullback-Leibler divergence. Note, however, that the formulation and conditions given above are comparable to those found in the well-specified situation (see *e.g.* Van der Vaart (1998) [91], chapter 10). Comparison to conditions (A1)–(A11) of Bunke and Milhaud (1998) [20] is also appropriate (see theorem 4.1 therein).

The following lemma collects some other important consequences of the conditions posed for lemma 2.2, expressed as properties of the score-function $\dot{\ell}_{\theta^*}$. (Note that the Lipschitz condition (2.34) which is slightly weaker than (2.12), is in fact sufficient in the proof.)

Lemma 2.3. *Under the conditions of P_0-almost-sure differentiability and Lipschitz continuity for $\theta \mapsto \log p_\theta(X)$ at θ^* (as in lemma 2.2), the score function is bounded as follows:*

$$\|\dot{\ell}_{\theta^*}(X)\| \leq m_{\theta^*}(X), \quad (P_0 - a.s.). \tag{2.16}$$

Furthermore, we have:

$$P_0\dot{\ell}_{\theta^*} = \partial_\theta\big[P_0\log p_\theta\big]_{\theta=\theta^*} = 0. \tag{2.17}$$

Proof P_0-almost-sure differentiability implies:

$$\Big|\log\frac{p_\theta}{p_{\theta^*}}\Big| = \big|(\theta-\theta^*)^T\dot{\ell}_{\theta^*} + o(\|\theta-\theta^*\|)\big| \leq m_{\theta^*}\|\theta-\theta^*\|, \quad (P_0 - a.s.),$$

[1] it is not unthinkable that a misspecified variation on Hellinger differentiability can be formulated and used to this end (by comparison, see theorem 7.2 in Van der Vaart (1998) [91]).

in the limit $\theta \to \theta^*$. Using the triangle-inequality, we derive (for P_0-almost-all X):

$$\Big| \dot{\ell}_{\theta^*}(X)^T \frac{(\theta - \theta^*)}{\|\theta - \theta^*\|} \Big| \leq m_{\theta^*}(X) + o(1),$$

Fix X, let $\epsilon > 0$ be given and choose $\eta = \epsilon m_{\theta^*}$. There exists a $\delta > 0$ such that for all θ with $\|\theta - \theta^*\| < \delta$, the absolute value of the $o(1)$-term in the above falls below ϵm_{θ^*}. Then the *r.h.s.* of the above display is bounded by $(1+\epsilon)m_{\theta^*}$. Choosing θ on a sphere of radius $\frac{1}{2}\delta$ around θ^* and taking the supremum, we find that $\|\dot{\ell}_{\theta^*}\| \leq (1+\epsilon)m_{\theta^*}$, P_0-almost-surely. Since this holds for all $\epsilon > 0$, we conclude that $\|\dot{\ell}_{\theta^*}\| \leq m_{\theta^*}$. To prove the second assertion, let $(\delta_n)_{n\geq 1}$ be a sequence such that $\delta_n \to 0$. Then:

$$\partial_\theta \big[P_0 \log p_\theta\big]_{\theta=\theta^*} = \lim_{n\to\infty} P_0\Big(\frac{1}{\delta_n} \log \frac{p_{\theta^*+\delta_n}}{p_{\theta^*}}\Big)$$

For large enough n, the modulus of the differential quotient on the *r.h.s.* is dominated by m_{θ^*}, $(P_0 - a.s.)$, which is P_0-integrable by assumption. This implies

$$\partial_\theta \big[P_0 \log p_\theta\big]_{\theta=\theta^*} = P_0 \dot{\ell}_{\theta^*},$$

by dominated convergence. The Kullback-Leibler divergence with respect to P_0 is differentiable at θ^* and since θ^* satisfies (2.2), we conclude that the previous display equals zero. □

2.3 Rate of convergence

In a Bayesian context, the rate of convergence is determined by the maximal speed at which balls around the point of convergence can be shrunk to radius zero while still capturing a posterior mass that converges to one asymptotically. Current interest lies in the fact that the formulation of the Bernstein-Von-Mises theorem as given in the previous section has condition (2.6). That requirement prescribes $\sqrt{n}$-rate, since for given sequence of radii $M_n \to \infty$:

$$\Pi_n\big(\, H \in K_n \;\big|\; X_1, \ldots, X_n \,\big) = \Pi_n\big(\, \|\underline{\theta} - \theta^*\| \leq M_n/\sqrt{n} \;\big|\; X_1, \ldots, X_n \,\big)$$

A convenient way of establishing the above is through the condition that suitable test sequences exist[2]. As has been shown in a well-specified context in Ghosal *et al.* (2000) [39] and under misspecification in Kleijn and Van der Vaart (2003) [57], the most important requirement for convergence of the posterior at a certain rate is the existence of a test-sequence that separates the point of convergence from the complements of balls shrinking at said rate.

This is also the approach we follow here: we show that the sequence of posterior probabilities in the above display converges to zero in P_0-probability if a test sequence exists that is

[2] Another condition ensures that sufficient prior mass is present in certain Kullback-Leibler neighbourhoods of the point of convergence. In the (parametric) case at hand, quite mild conditions on the prior suffice (see 2.2.1), but in non-parametric models this changes.

suitable in the sense given above (see the proof of theorem 2.2). However, under the regularity conditions that were formulated to establish local asymptotic normality under misspecification in the previous section, more can be said: not complements of shrinking balls, but fixed alternatives are to be suitably testable against P_0, thus relaxing the testing condition considerably. Locally, the construction relies on score-tests to separate the point of convergence from complements of neighbourhoods shrinking at rate $1/\sqrt{n}$, using Bernstein's inequality to obtain exponential power. The tests for fixed alternatives are used to extend those local tests to the full model.

In this section we prove that a prior mass condition and suitable test sequences suffice to prove convergence at the rate required for the Bernstein-Von-Mises theorem as formulated in section 2.2. The theorem that begins the next subsection summarizes the conclusion.

2.3.1 Posterior rate of convergence

With use of theorem 2.3, we formulate a theorem that ensures $\sqrt{n}$-rate of convergence for the posterior distributions of smooth, testable models with sufficient prior mass around the point of convergence. The testability condition is formulated using measures Q_θ, which are defined as follows:

$$Q_\theta(A) = P_0\Big(\frac{p_\theta}{p_{\theta^*}} 1_A\Big)$$

for all $A \in \mathscr{A}$ and all $\theta \in \Theta$. Note that Q_θ is not necessarily a finite measure, that all Q_θ are dominated by P_0 and that $Q_{\theta^*} = P_0$. Also note that if the model is well specified, $P_{\theta^*} = P_0$ which implies that $Q_\theta = P_\theta$ for all θ. Therefore the use of Q_θ instead of P_θ to formulate the testing condition is relevant only in the misspecified situation (see Kleijn and Van der Vaart (2003) [57] for more on this subject).

Theorem 2.2. *Assume that the model satisfies the smoothness conditions of lemma 2.2 and that the prior has the properties described in the second paragraph of subsection 2.2.1. Furthermore, assume that for every $\epsilon > 0$ there exists a sequence of tests $(\phi_n)_{n\geq 1}$ such that:*

$$P_0^n \phi_n \to 0, \qquad \sup_{\{\theta:\|\theta-\theta^*\|\geq\epsilon\}} Q_\theta^n(1-\phi_n) \to 0.$$

Then the posterior converges at rate $1/\sqrt{n}$, i.e. for every sequence $(M_n)_{n\geq 1}$, $M_n \to \infty$:

$$\Pi\big(\,\theta \in \Theta : \|\theta - \theta^*\| \geq M_n/\sqrt{n} \;\big|\; X_1, X_2, \ldots, X_n\,\big) \xrightarrow{P_0} 0.$$

Proof Let $(M_n)_{n\geq 1}$ be given, define the sequence $(\epsilon_n)_{n\geq 1}$ by $\epsilon_n = M_n/\sqrt{n}$. According to lemma 2.3 there exists a sequence of tests $(\omega_n)_{n\geq 1}$ and a constant $D > 0$, such that (2.23) holds. We use these tests to split the P_0^n-expectation of the posterior measure as follows:

$$\begin{aligned} &P_0^n \Pi\big(\theta : \|\theta - \theta^*\| \geq \epsilon_n \;\big|\; X_1, X_2, \ldots, X_n\big) \\ &\qquad \leq P_0^n \omega_n + P_0^n(1-\omega_n)\Pi\big(\theta : \|\theta - \theta^*\| \geq \epsilon_n \;\big|\; X_1, X_2, \ldots, X_n\big). \end{aligned}$$

Note that the first term is of order $o(1)$ as $n \to \infty$. Given a constant $\epsilon > 0$ (to be specified later), the second term is decomposed as:

$$\begin{aligned} P_0^n(1-\omega_n)\Pi\big(\theta : \|\theta-\theta^*\| \geq \epsilon_n \,\big|\, X_1, X_2, \ldots, X_n\big) \\ = P_0^n(1-\omega_n)\Pi\big(\theta : \|\theta-\theta^*\| \geq \epsilon \,\big|\, X_1, X_2, \ldots, X_n\big) \\ + P_0^n(1-\omega_n)\Pi\big(\theta : \epsilon_n \leq \|\theta-\theta^*\| < \epsilon \,\big|\, X_1, X_2, \ldots, X_n\big). \end{aligned} \tag{2.18}$$

Given two constants $M, M' > 0$ (also to be specified at a later stage), we define the sequences $(a_n)_{n\geq 1}$, $a_n = M\sqrt{\log n/n}$ and $(b_n)_{n\geq 1}$, $b_n = M'\epsilon_n$. Based on a_n and b_n, we define two sequences of events:

$$\Xi_n = \Big\{ \int_\Theta \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\, d\Pi(\theta) \leq \Pi\big(B(a_n, \theta^*; P_0)\big) e^{-na_n^2(1+C)} \Big\},$$

$$\Omega_n = \Big\{ \int_\Theta \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\, d\Pi(\theta) \leq \Pi\big(B(b_n, \theta^*; P_0)\big) e^{-nb_n^2(1+C)} \Big\}.$$

The sequence $(\Xi_n)_{n\geq 1}$ is used to split the first term on the *r.h.s.* of (2.18) and estimate it as follows:

$$\begin{aligned} P_0^n(1-\omega_n)\Pi\big(\theta : \|\theta-\theta^*\| \geq \epsilon \,\big|\, X_1, X_2, \ldots, X_n\big) \\ \leq P_0(\Xi_n) + P_0^n(1-\omega_n)\, 1_{\Omega\setminus\Xi_n}\, \Pi\big(\theta : \|\theta-\theta^*\| \geq \epsilon \,\big|\, X_1, X_2, \ldots, X_n\big). \end{aligned}$$

According to lemma 2.4, the first term is of order $o(1)$ as $n \to \infty$. The second term is estimated further with the use of lemmas 2.4 and 2.5 (for some $C > 0$):

$$\begin{aligned} P_0^n(1-\omega_n)\, 1_{\Omega\setminus\Xi_n}\, \Pi\big(\theta : \|\theta-\theta^*\| \geq \epsilon \,\big|\, X_1, X_2, \ldots, X_n\big) \\ \leq \frac{e^{na_n^2(1+C)}}{\Pi\big(B(a_n, \theta^*; P_0)\big)} \int_{\{\theta : \|\theta-\theta^*\| \geq \epsilon\}} P_0^n\Big(\prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)(1-\omega_n)\Big)\, d\Pi(\theta) \\ = \frac{e^{na_n^2(1+C)}}{\Pi\big(B(a_n, \theta^*; P_0)\big)} \int_{\{\theta : \|\theta-\theta^*\| \geq \epsilon\}} Q_\theta^n(1-\omega_n)\, d\Pi(\theta) \\ \leq \frac{e^{n(a_n^2(1+C) - D\epsilon^2)}}{\Pi\big(B(a_n, \theta^*; P_0)\big)} \Pi\big(\,\theta : \|\theta-\theta^*\| \geq \epsilon\,\big). \end{aligned}$$

In the last step, we make use of the uniform bound on the power of the test function, as given in theorem 2.3. Since $a_n \to 0$, $a_n^2(1+C) \leq \frac{1}{2}D\epsilon^2$ for large enough n and therefore $a_n^2(1+C) - D\epsilon^2 \leq -a_n^2(1+C)$, which we use as follows:

$$\begin{aligned} P_0^n(1-\omega_n)\, 1_{\Omega\setminus\Xi_n}\, \Pi\big(\theta : \|\theta-\theta^*\| \geq \epsilon \,\big|\, X_1, X_2, \ldots, X_n\big) \leq \frac{e^{n(a_n^2(1+C) - D\epsilon^2)}}{\Pi\big(B(a_n, \theta^*; P_0)\big)} \\ \leq K^{-1} e^{-na_n^2(1+C)} (a_n)^{-d} \leq \frac{M^{d/2}}{K} (\log n)^{-d/2} n^{-M(1+C)+\frac{d}{2}}, \end{aligned}$$

for large enough n, using (2.20). A suitably large choice of M then ensures that the expression on the *l.h.s.* in the previous display is of order $o(1)$ as $n \to \infty$.

The sequence $(\Omega_n)_{n\geq 1}$ is used to split the second term on the *r.h.s.* of (2.18) after which we estimate it in a similar manner. Again the term that derives from 1_{Ω_n} is of order $o(1)$, and

$$\begin{aligned}
&P_0^n(1-\omega_n)\,1_{\Omega\setminus\Xi_n}\,\Pi\big(\theta:\epsilon_n\leq\|\theta-\theta^*\|<\epsilon\;\big|\;X_1,X_2,\ldots,X_n\big)\\
&\quad\leq\frac{e^{n\epsilon_n^2(1+C)}}{\Pi\big(B(\epsilon_n,\theta^*;P_0)\big)}\int_{\{\theta:\epsilon_n\leq\|\theta-\theta^*\|<\epsilon\}}P_0^n\Big(\prod_{i=1}^n\frac{p_\theta}{p_{\theta^*}}(X_i)(1-\omega_n)\Big)\,d\Pi(\theta)\\
&\quad=\frac{e^{n\epsilon_n^2(1+C)}}{\Pi\big(B(\epsilon_n,\theta^*;P_0)\big)}\int_{\{\theta:\epsilon_n\leq\|\theta-\theta^*\|<\epsilon\}}Q_\theta^n(1-\omega_n)\,d\Pi(\theta)\\
&\quad=\frac{e^{n\epsilon_n^2(1+C)}}{\Pi\big(B(\epsilon_n,\theta^*;P_0)\big)}\sum_{j=1}^J\int_{A_{n,j}}Q_\theta^n(1-\omega_n)\,d\Pi(\theta).
\end{aligned}$$

where we have split the domain of integration into spherical shells $A_{n,j}$, $(1\leq j\leq J$, with J the smallest integer such that $(J+1)\epsilon_n>\epsilon)$, of width ϵ_n defined as follows:

$$A_{n,j}=\big\{\,\theta:j\epsilon_n\leq\|\theta-\theta^*\|\leq\big((j+1)\epsilon_n\big)\wedge\epsilon\,\big\}.$$

Applying theorem 2.3 to each of the shells separately, we obtain:

$$\begin{aligned}
&P_0^n(1-\omega_n)\,1_{\Omega\setminus\Xi_n}\,\Pi\big(\theta:\epsilon_n\leq\|\theta-\theta^*\|<\epsilon\;\big|\;X_1,X_2,\ldots,X_n\big)\\
&\quad=\sum_{j=1}^J e^{n\epsilon_n^2(1+C)}\sup_{\theta\in A_{n,j}}Q_\theta^n(1-\omega_n)\;\frac{\Pi(A_{n,j})}{\Pi\big(B(\epsilon_n,\theta^*;P_0)\big)}\\
&\quad\leq\sum_{j=1}^J e^{n\epsilon_n^2(1+C)-nDj^2\epsilon_n^2}\frac{\Pi\big\{\,\theta:\|\theta-\theta^*\|\leq(j+1)\epsilon_n\,\big\}}{\Pi\big(B(\epsilon_n,\theta^*;P_0)\big)}
\end{aligned}$$

For a small enough choice of ϵ and large enough n, the sets $\big\{\,\theta:\|\theta-\theta^*\|\leq(j+1)\epsilon_n\,\big\}$ all fall within the neighbourhood U of θ^* on which the prior density π is continuous. Hence π is uniformly bounded by a constant $R>0$ and we see that:

$$\Pi\big\{\,\theta:\|\theta-\theta^*\|\leq(j+1)\epsilon_n\,\big\}\leq RV_d(j+1)^d\epsilon_n^d.$$

Combining this with (2.20), we see that there exists a constant $K'>0$ such that for large enough n (and with the choice $M'<D^2/2(1+C)$):

$$\begin{aligned}
&P_0^n(1-\omega_n)\,1_{\Omega\setminus\Xi_n}\,\Pi\big(\theta:\epsilon_n\leq\|\theta-\theta^*\|<\epsilon\;\big|\;X_1,\ldots,X_n\big)\\
&\quad\leq K'\sum_{j=1}^J e^{n\epsilon_n^2(1+C)-nDj^2\epsilon_n^2}(j+1)^d\\
&\quad\leq K'e^{-nM'\epsilon_n^2(1+C)}\sum_{j=1}^\infty(j+1)^d e^{-nD(j^2-1)\epsilon_n^2}.
\end{aligned}$$

The series is convergent and we conclude that this term is also of order $o(1)$ as $n\to\infty$. □

In the above proof, lower bounds in probability on the denominators of posterior probabilities (*c.f.* (2.3)) are needed. The following lemma provides the required bound, expressing it in terms of the prior mass of Kullback-Leibler neighbourhoods of θ^* of the form:

$$B(\epsilon, \theta^*; P_0) = \Big\{\theta \in \Theta : -P_0 \log \frac{p_\theta}{p_{\theta^*}} \leq \epsilon^2,\, P_0\Big(\log \frac{p_\theta}{p_{\theta^*}}\Big)^2 \leq \epsilon^2\Big\}. \tag{2.19}$$

(for some $\epsilon > 0$).

Lemma 2.4. *For given $\epsilon > 0$ and $\theta^* \in \Theta$ such that $P_0 \log(p_0/p_{\theta^*}) < \infty$ define $B(\epsilon, \theta^*; P_0)$ by (2.19). Then for every $C > 0$ and probability measure Π on Θ:*

$$P_0^n\Big(\int_\Theta \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\, d\Pi(\theta) \leq \Pi\big(B(\epsilon, \theta^*; P_0)\big) e^{-n\epsilon^2(1+C)}\Big) \leq \frac{1}{C^2 n \epsilon^2}.$$

Proof This lemma can also be found as lemma 7.1 in Kleijn and Van der Vaart (2003) [57]. The proof is analogous to that of lemma 8.1 in Ghosal *et al.* (2000) [39]. □

Moreover, the prior mass of the Kullback-Leibler neighbourhoods $\Pi\big(B(\epsilon, \theta^*; P_0)\big)$ can be lower-bounded if we make the regularity assumptions for the model used in section 2.2 and the assumption the prior has a Lebesgue density that is well-behaved at θ^*.

Lemma 2.5. *Under the smoothness conditions of lemma 2.2 and assuming that the prior density π is continuous and strictly positive in θ^*, there exists a constant $K > 0$ such that the prior mass of the Kullback-Leibler neighbourhoods $B(\epsilon, \theta^*; P_0)$ satisfies:*

$$\Pi\big(B(\epsilon, \theta^*; P_0)\big) \geq K\epsilon^d. \tag{2.20}$$

for small enough $\epsilon > 0$.

Proof As a result of the smoothness conditions, we have, for some constants $d_1, d_2 > 0$ and small enough $\|\theta - \theta^*\|$:

$$-P_0 \log \frac{p_\theta}{p_{\theta^*}} \leq d_1 \|\theta - \theta^*\|^2, \qquad P_0\big(\log \frac{p_\theta}{p_{\theta^*}}\big)^2 \leq d_2 \|\theta - \theta^*\|^2.$$

Defining $d = (1/d_1 \wedge 1/d_2)^{1/2}$, this implies that for small enough $\epsilon > 0$,

$$\big\{\theta \in \Theta : \|\theta - \theta^*\| \leq d\epsilon\big\} \subset B(\epsilon, \theta^*; P_0). \tag{2.21}$$

Since the Lebesgue-density π of the prior is continuous and strictly positive in θ^*, we see that there exists a $\delta' > 0$ such that for all $0 < \delta \leq \delta'$:

$$\Pi\big(\theta \in \Theta : \|\theta - \theta^*\| \leq \delta\big) \geq \tfrac{1}{2} V_d \pi(\theta^*)\, \delta^d > 0, \tag{2.22}$$

where V_d is the Lebesgue-volume of the d-dimensional ball of unit radius. Hence, for small enough ϵ, $d\epsilon \leq \delta'$ and we obtain (2.20) upon combination of (2.21) and (2.22). □

2.3.2 Suitable test sequences

In this subsection we prove that the existence of test sequences (under misspecification) of uniform exponential power for complements of shrinking balls around θ^* versus P_0 (as needed in the proof of theorem 2.2), is guaranteed whenever asymptotically consistent test-sequences exist for complements of *fixed* balls around θ^* versus P_0 and the conditions of lemmas 2.2 and 2.7 are met. The following theorem is inspired by lemma 10.3 in Van der Vaart (1998) [91].

Theorem 2.3. *Assume that the conditions of lemma 2.2 are satisfied, where in addition, it is required that $P_0(p_\theta/p_{\theta^*}) < \infty$ for all θ in a neighbourhood of θ^* and $P_0(e^{sm_{\theta^*}}) < \infty$ for some $s > 0$. Furthermore, suppose that for every $\epsilon > 0$ there exists a sequence of test functions $(\phi_n)_{n\geq 1}$, such that:*

$$P_0^n \phi_n \to 0, \qquad \sup_{\{\theta:\|\theta-\theta^*\|\geq\epsilon\}} Q_\theta^n(1-\phi_n) \to 0.$$

Then for every sequence $(M_n)_{n\geq 1}$ such that $M_n \to \infty$ there exists a sequence of tests $(\omega_n)_{n\geq 1}$ such that for some constants $D > 0$, $\epsilon > 0$ and large enough n:

$$P_0^n \omega_n \to 0, \quad Q_\theta^n(1-\omega_n) \leq e^{-nD(\|\theta-\theta^*\|^2\wedge\epsilon^2)}, \tag{2.23}$$

for all $\theta \in \Theta$ such that $\|\theta-\theta^*\| \geq M_n/\sqrt{n}$.

Proof Let $(M_n)_{n\geq 1}$ be given. We construct two sequences of tests: one sequence to test P_0 versus $\{Q_\theta : \theta \in \Theta_1\}$ with $\Theta_1 = \{\theta \in \Theta : M_n/\sqrt{n} \leq \|\theta-\theta^*\| \leq \epsilon\}$, and the other to test P_0 versus $\{Q_\theta : \theta \in \Theta_2\}$ with $\Theta_2 = \{\theta : \|\theta-\theta^*\| > \epsilon\}$, both uniformly with exponential power (for a suitable choice of ϵ). We combine these sequences to test P_0 versus $\{Q_\theta : \|\theta-\theta^*\| \geq M_n/\sqrt{n}\}$ uniformly with exponential power.

For the construction of the first sequence, a constant $L > 0$ is chosen to truncate the score-function coordinate-wise (*i.e.* for all $1 \leq k \leq d$, $(\dot{\ell}^L_{\theta^*})_k = 0$ if $|(\dot{\ell}_{\theta^*})_k| \geq L$ and $(\dot{\ell}^L_{\theta^*})_k = (\dot{\ell}_{\theta^*})_k$ otherwise) and we define:

$$\omega_{1,n} = 1\big\{\|(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*}\| > \sqrt{M_n/n}\big\},$$

Because the function $\dot{\ell}_{\theta^*}$ is square integrable, we can ensure that the matrices $P_0(\dot{\ell}_{\theta^*}\dot{\ell}^T_{\theta^*})$, $P_0(\dot{\ell}_{\theta^*}(\dot{\ell}^L_{\theta^*})^T)$ and $P_0(\dot{\ell}^L_{\theta^*}(\dot{\ell}^L_{\theta^*})^T)$ are arbitrarily close (for instance in operator norm) by a sufficiently large choice for the constant L. We fix such an L throughout the proof.

By the central limit theorem $P_0^n\omega_{1,n} = P_0^n\big(\,\|\sqrt{n}(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*}\|^2 > M_n\,\big) \to 0$. Turning to $Q_\theta^n(1-\omega_{1,n})$ for $\theta \in \Theta_1$, we note that for all θ:

$$\begin{aligned} Q_\theta^n\Big(\|(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*}\| \leq \sqrt{M_n/n}\Big) &= Q_\theta^n\Big(\sup_{v\in S} v^T(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*} \leq \sqrt{M_n/n}\Big) \\ &= Q_\theta^n\Big(\bigcap_{v\in S}\big\{v^T(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*} \leq \sqrt{M_n/n}\big\}\Big) \leq \inf_{v\in S} Q_\theta^n\Big(v^T(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*} \leq \sqrt{M_n/n}\Big), \end{aligned}$$

where S is the sphere of unity in $\mathbb{R}^d$. With the choice $v = (\theta - \theta^*)/\|\theta - \theta^*\|$ as an upper bound for the *r.h.s.* in the above display, we note that:

$$\begin{aligned} &Q_\theta^n\Big((\theta-\theta^*)^T(\mathbb{P}_n - P_0)\dot\ell^L_{\theta^*} \le \sqrt{M_n/n}\|\theta-\theta^*\|\Big) \\ &\quad = Q_\theta^n\Big((\theta^*-\theta)^T(\mathbb{P}_n - \tilde Q_\theta)\dot\ell^L_{\theta^*} \ge (\theta-\theta^*)^T(\tilde Q_\theta - \tilde Q_{\theta^*})\dot\ell^L_{\theta^*} - \sqrt{M_n/n}\|\theta-\theta^*\|\Big). \end{aligned} \tag{2.24}$$

where we have used the notation (for all $\theta \in \Theta_1$ with small enough $\epsilon > 0$) $\tilde Q_\theta = \|Q_\theta\|^{-1}Q_\theta$ and the fact that $P_0 = Q_{\theta^*} = \tilde Q_{\theta^*}$. By straightforward manipulation, we find:

$$\begin{aligned} &(\theta-\theta^*)^T\big(\tilde Q_\theta - \tilde Q_{\theta^*}\big)\dot\ell^L_{\theta^*} \\ &\quad = \frac{1}{P_0(p_\theta/p_{\theta^*})}(\theta-\theta^*)^T\Big(P_0\big((p_\theta/p_{\theta^*}-1)\dot\ell^L_{\theta^*}\big) + \big(1 - P_0(p_\theta/p_{\theta^*})\big)P_0\dot\ell^L_{\theta^*}\Big). \end{aligned} \tag{2.25}$$

In view of lemma 2.7 and conditions (2.12), (2.13), $(P_0(p_\theta/p_{\theta^*}) - 1)$ is of order $O(\|\theta-\theta^*\|^2)$ as $(\theta \to \theta^*)$, which means that if we approximate the above display up to order $o(\|\theta-\theta^*\|^2)$, we can limit attention on the *r.h.s.* to the first term in the last factor and equate the first factor to 1. Furthermore, using the differentiability of $\theta \mapsto \log(p_\theta/p_{\theta^*})$, condition (2.12) and lemma 2.7, we see that:

$$\begin{aligned} &P_0\Big\|\Big(\frac{p_\theta}{p_{\theta^*}} - 1 - (\theta-\theta^*)^T\dot\ell_{\theta^*}\Big)\dot\ell^L_{\theta^*}\Big\| \\ &\quad \le P_0\Big\|\Big(\frac{p_\theta}{p_{\theta^*}} - 1 - \log\frac{p_\theta}{p_{\theta^*}}\Big)\dot\ell^L_{\theta^*}\Big\| + P_0\Big\|\Big(\log\frac{p_\theta}{p_{\theta^*}} - (\theta-\theta^*)^T\dot\ell_{\theta^*}\Big)\dot\ell^L_{\theta^*}\Big\| = o\big(\|\theta-\theta^*\|\big). \end{aligned}$$

Also note that since $M_n \to \infty$ and for all $\theta \in \Theta_1$, $\|\theta-\theta^*\| \ge M_n/\sqrt{n}$,

$$-\|\theta-\theta^*\|\sqrt{M_n/n} \ge -\|\theta-\theta^*\|^2\frac{1}{\sqrt{M_n}}.$$

Summarizing the above and combining with the remark made at the beginning of the proof concerning the choice of L, we find that for every $\delta > 0$, there exist choices of $\epsilon > 0$, $L > 0$ and $N \ge 1$ such that for all $n \ge N$ and all θ in Θ_1:

$$(\theta-\theta^*)^T\big(\tilde Q_\theta - \tilde Q_{\theta^*}\big)\dot\ell^L_{\theta^*} - \sqrt{M_n/n}\|\theta-\theta^*\| \ge (\theta-\theta^*)^T P_0\big(\dot\ell_{\theta^*}\dot\ell^T_{\theta^*}\big)(\theta-\theta^*) - \delta\|\theta-\theta^*\|^2.$$

We denote $\Delta(\theta) = (\theta-\theta^*)^T P_0(\dot\ell_{\theta^*}\dot\ell^T_{\theta^*})(\theta-\theta^*)$ and we assume that $\Delta(\theta) > 0$ (we discuss the case $\Delta(\theta) = 0$ separately at a later stage). Choosing $c > 0$ to be the smallest non-zero eigenvalue of $P_0(\dot\ell_{\theta^*}\dot\ell^T_{\theta^*})$ (note that this matrix is positive) we see that $-\delta\|\theta-\theta^*\|^2 \ge -\delta/c\,\Delta(\theta)$. Hence there exists a constant $r(\delta)$ (depending only on the matrix $P_0(\dot\ell_{\theta^*}\dot\ell^T_{\theta^*})$ and with the property that $r(\delta) \to 1$ if $\delta \to 0$) such that:

$$Q_\theta^n(1-\omega_{1,n}) \le Q_\theta^n\Big((\theta^*-\theta)^T(\mathbb{P}_n - \tilde Q_\theta)\dot\ell^L_{\theta^*} \ge r(\delta)\Delta(\theta)\Big),$$

for small enough ϵ, large enough L and large enough n, demonstrating that the type-2 error is bounded above by the (unnormalized) tail probability $Q_\theta^n(\bar W_n \ge r(\delta)\Delta(\theta))$ of the mean of the variables $(1 \le i \le n)$:

$$W_i = (\theta^*-\theta)^T(\dot\ell^L_{\theta^*}(X_i) - \tilde Q_\theta\dot\ell^L_{\theta^*}),$$

so that $\tilde{Q}_\theta W_i = 0$. The random variables W_i are independent and bounded since:

$$|W_i| \leq \|\theta - \theta^*\|\big(\|\dot{\ell}^L_{\theta^*}(X_i)\| + \|\tilde{Q}_\theta \dot{\ell}^L_{\theta^*}\|\big) \leq 2L\sqrt{d}\|\theta - \theta^*\|.$$

The variance of W_i under $\tilde{Q}_\theta$ is expressed as follows:

$$\mathrm{Var}_{\tilde{Q}_\theta} W_i = (\theta - \theta^*)^T \Big[\tilde{Q}_\theta\big(\dot{\ell}^L_{\theta^*}(\dot{\ell}^L_{\theta^*})^T\big) - \tilde{Q}_\theta \dot{\ell}^L_{\theta^*} \tilde{Q}_\theta (\dot{\ell}^L_{\theta^*})^T\Big](\theta - \theta^*)$$

Referring to the argument following (2.25) and using that $P_0 \dot{\ell}_{\theta^*} = 0$ (see lemma 2.3), the above can be estimated like before, with the result that there exists a constant $s(\delta)$ (depending only on (the largest eigenvalue of) the matrix $P_0(\dot{\ell}_{\theta^*}\dot{\ell}^T_{\theta^*})$ and with the property that $s(\delta) \to 1$ as $\delta \to 0$) such that:

$$\mathrm{Var}_{\tilde{Q}_\theta}(W_i) \leq s(\delta)\Delta(\theta),$$

for small enough ϵ and large enough L. We apply Bernstein's inequality (see, for instance, Pollard (1984) [74], pp. 192–193) to obtain:

$$\begin{aligned} Q^n_\theta(1 - \omega_{1,n}) &= \|Q_\theta\|^n \, \tilde{Q}^n_\theta\big(W_1 + \ldots + W_n \geq nr(\delta)\Delta(\theta)\big) \\ &\leq \|Q_\theta\|^n \exp\Big(-\frac{1}{2}\frac{r(\delta)^2\, n\, \Delta(\theta)}{s(\delta) + \frac{3}{2}L\sqrt{d}\|\theta - \theta^*\| r(\delta)}\Big). \end{aligned} \tag{2.26}$$

The factor $t(\delta) = r(\delta)^2(s(\delta) + \frac{3}{2}L\sqrt{d}\|\theta - \theta^*\| r(\delta))^{-1}$ lies arbitrarily close to 1 for sufficiently small choices of δ and ϵ. As for the n-th power of the norm of Q_θ, we use lemma 2.7, (2.12) and (2.13) to estimate the norm of Q_θ as follows:

$$\begin{aligned} \|Q_\theta\| &= 1 + P_0 \log \frac{p_\theta}{p_{\theta^*}} + \tfrac{1}{2} P_0\Big(\log \frac{p_\theta}{p_{\theta^*}}\Big)^2 + o(\|\theta - \theta^*\|^2) \\ &\leq 1 + P_0 \log \frac{p_\theta}{p_{\theta^*}} + \tfrac{1}{2}(\theta - \theta^*)^T P_0\big(\dot{\ell}_{\theta^*}\dot{\ell}^T_{\theta^*}\big)(\theta - \theta^*) + o(\|\theta - \theta^*\|^2) \\ &\leq 1 - \tfrac{1}{2}(\theta - \theta^*)^T V_{\theta^*}(\theta - \theta^*) + \tfrac{1}{2}u(\delta)\Delta(\theta) \end{aligned} \tag{2.27}$$

for some constant $u(\delta)$ such that $u(\delta) \to 1$ if $\delta \to 0$. Because $1 + x \leq e^x$ for all $x \in \mathbb{R}$, we obtain, for sufficiently small $\|\theta - \theta^*\|$:

$$Q^n_\theta(1 - \omega_{1,n}) \leq \exp\Big(-\frac{n}{2}(\theta - \theta^*)^T V_{\theta^*}(\theta - \theta^*) + \frac{n}{2}\big(u(\delta) - t(\delta)\big)\Delta(\theta)\Big). \tag{2.28}$$

Note that $u(\delta) - t(\delta) \to 0$ as $\delta \to 0$ and $\Delta(\theta)$ is upper bounded by a multiple of $\|\theta - \theta^*\|^2$. Since V_{θ^*} is assumed to be invertible, we conclude that there exists a constant $C > 0$ such that for large enough L, small enough $\epsilon > 0$ and large enough n:

$$Q^n_\theta(1 - \omega_{1,n}) \leq e^{-Cn\|\theta - \theta^*\|^2}. \tag{2.29}$$

Coming back to the assumption $\Delta(\theta) > 0$, we note the following: if θ is such that $\Delta(\theta) = 0$, we can omit the discussion that led to (2.26) and immediately estimate $Q^n_\theta(1 - \omega_{1,n})$ by the n-th power of the norm of Q_θ. In that case, the term of order $o(\|\theta - \theta^*\|^2)$ in (2.27) is absorbed

by means of a constant $v(\delta)$ (depending only on the matrix V_{θ^*} and with the property that $v(\delta) \to 1$ as $\delta \to 0$) when we replace (2.28) by:

$$Q_\theta^n(1-\omega_{1,n}) \leq \|Q_\theta\|^n \leq \exp\Big(-\frac{n}{2}v(\delta)(\theta-\theta^*)^T V_{\theta^*}(\theta-\theta^*)\Big),$$

leading to (2.29) by the same argument.

As for the range $\|\theta-\theta^*\| > \epsilon$, an asymptotically consistent test-sequence of P_0 versus Q_θ exists by assumption, what remains is the exponential power; the proof of lemma 2.6 demonstrates the existence of a sequence of tests $(\omega_{2,n})_{n\geq 1}$ such that (2.30) holds. The sequence $(\psi_n)_{n\geq 1}$ is defined as the maximum of the two sequences defined above: $\psi_n = \omega_{1,n} \vee \omega_{2,n}$ for all $n \geq 1$, in which case $P_0^n\psi_n \leq P_0^n\omega_{1,n} + P_0^n\omega_{2,n} \to 0$ and:

$$\begin{aligned}\sup_{\theta\in A_n} Q_\theta^n(1-\psi_n) &= \sup_{\theta\in\Theta_1} Q_\theta^n(1-\psi_n) \vee \sup_{\theta\in\Theta_2} Q_\theta^n(1-\psi_n) \\ &\leq \sup_{\theta\in\Theta_1} Q_\theta^n(1-\omega_{1,n}) \vee \sup_{\theta\in\Theta_2} Q_\theta^n(1-\omega_{2,n}).\end{aligned}$$

Combination of the bounds found in (2.29) and (2.30) and a suitable choice for the constant $D > 0$ lead to (2.23). □

The following lemma shows that for a sequence of tests that separates P_0 from a fixed model subset V, there exists a exponentially powerful version without further conditions. Note that this lemma holds in non-parametric and parametric situations alike.

Lemma 2.6. *Suppose that for given measurable subset V of Θ, there exists a sequence of tests $(\phi_n)_{n\geq 1}$ such that:*

$$P_0^n\phi_n \to 0, \qquad \sup_{\theta\in V} Q_\theta^n(1-\phi_n) \to 0.$$

Then there exists a sequence of tests $(\omega_n)_{n\geq 1}$ and strictly positive constants C, D such that:

$$P_0^n\omega_n \leq e^{-nC}, \qquad \sup_{\theta\in V} Q_\theta^n(1-\omega_n) \leq e^{-nD} \tag{2.30}$$

Proof For given $0 < \zeta < 1$, we split the model subset V in two disjoint parts V_1 and V_2 defined by:

$$V_1 = \big\{\theta \in V : \|Q_\theta\| \geq 1-\zeta\big\}, \qquad V_2 = \big\{\theta \in V : \|Q_\theta\| < 1-\zeta\big\}$$

Note that for every test-sequence $(\omega_n)_{n\geq 1}$,

$$\begin{aligned}\sup_{\theta\in V} Q_\theta^n(1-\omega_n) &= \sup_{\theta\in V_1} Q_\theta^n(1-\omega_n) \vee \sup_{\theta\in V_2} Q_\theta^n(1-\omega_n) \\ &\leq \sup_{\theta\in V_1} Q_\theta^n(1-\omega_n) \vee (1-\zeta)^n.\end{aligned} \tag{2.31}$$

Let $\delta > 0$ be given. By assumption there exists an $N \geq 1$ such that for all $n \geq N+1$,

$$P_0^n\phi_n \leq \delta, \qquad \sup_{\theta\in V} Q_\theta^n(1-\phi_n) \leq \delta. \tag{2.32}$$

Every $n \geq N+1$ can be written as an m-fold multiple of N ($m \geq 1$) plus a remainder $1 \leq r \leq N$: $n = mN + r$. Given $n \geq N$, we divide the sample $X_1, X_2, \ldots, X_n$ into $(m-1)$ groups of N consecutive X's and a group of $N+r$ X's and apply ϕ_N to the first $(m-1)$ groups and ϕ_{N+r} to the last group, to obtain:

$$\begin{aligned} Y_{1,n} &= \phi_N\big(X_1, X_2, \ldots, X_N\big) \\ Y_{2,n} &= \phi_N\big(X_{N+1}, X_{N+2}, \ldots, X_{2N}\big) \\ &\vdots \\ Y_{m-1,n} &= \phi_N\big(X_{(m-2)N+1}, X_{(m-2)N+2}, \ldots, X_{(m-1)N}\big) \\ Y_{m,n} &= \phi_{N+r}\big(X_{(m-1)N+1}, X_{(m-1)N+2}, \ldots, X_{mN+r}\big) \end{aligned}$$

which are bounded, $0 \leq Y_{j,n} \leq 1$ for all $1 \leq j \leq m$ and $n \geq 1$. From that we define the following test-statistic:

$$\overline{Y}_{m,n} = \frac{1}{m}(Y_{1,n} + \ldots + Y_{m,n}).$$

and the test-function based on a critical value $\eta > 0$ to be chosen at a later stage:

$$\omega_n = 1\big\{\overline{Y}_{m,n} \geq \eta\big\}.$$

Using the first bound in (2.32), the P_0^n-expectation of the test-function can be bounded as follows:

$$\begin{aligned} P_0^n\omega_n &= P_0^n\big(Y_{1,n} + \ldots + Y_{m,n} \geq m\eta\big) \\ &= P_0^n\Big(Z_{1,n} + \ldots + Z_{m,n} \geq m\eta - \sum_{j=1}^{m-1} P_0^n Y_{j,n} - P_0^{N+r} Y_{m,n}\Big) \\ &\leq P_0^n\big(Z_{1,n} + \ldots + Z_{m,n} \geq m(\eta - \delta)\big) \end{aligned}$$

where $Z_{j,n} = Y_{j,n} - P_0^n Y_{j,n}$ for all $1 \leq j \leq m-1$ and $Z_{m,n} = Y_{m,n} - P_0^{N+r} Y_{m,n}$. Furthermore, the variables $Z_{j,n}$ are bounded $a_j \leq Z_{j,n} \leq b_j$ where $b_j - a_j = 1$. Imposing $\eta > \delta$ we may use Hoeffding's inequality to conclude that:

$$P_0^n\omega_n \leq e^{-2m(\eta-\delta)^2}. \tag{2.33}$$

A similar bound can be derived for $Q_\theta(1-\omega_n)$ as follows. First we note that:

$$\begin{aligned} Q_\theta^n(1-\omega_n) &= Q_\theta\big(\overline{Y}_{m,n} > \eta\big) \\ &= Q_\theta^n\big(Y_{1,n} + \ldots + Y_{m,n} < m\eta\big) \\ &= Q_\theta^n\Big(Z_{1,n} + \ldots + Z_{m,n} > -m\eta + \sum_{j=1}^{m-1} Q_\theta^N Y_{j,n} + Q_\theta^{N+r} Y_{m,n}\Big), \end{aligned}$$

where, in this case, we have used the following definitions for the variables $Z_{j,n}$:

$$Z_{j,n} = -Y_{j,n} + Q_\theta^N Y_{j,n}, \qquad Z_{m,n} = -Y_{m,n} + Q_\theta^{N+r} Y_{m,n},$$

for $1 \leq j \leq m-1$. We see that $a_j \leq Z_{j,n} \leq b_j$ with $b_j - a_j = 1$. Choosing $\zeta \leq 1 - (4\delta)^{1/N}$ (for small enough $\delta > 0$) and η between δ and 2δ, we see that for all $\theta \in V_1$:

$$\sum_{j=1}^{m-1} Q_\theta^N Y_{j,n} + Q_\theta^{N+r} Y_{m,n} - m\eta \geq m(\|Q_\theta\|^N - \delta - \eta) \geq m\big((1-\zeta)^N - 3\delta\big) \geq m\delta > 0.$$

based on the second bound in (2.32). This implies that Hoeffding's inequality (see, for instance, Pollard (1984) [74], theorem B.2) can be applied with the following result:

$$Q_\theta^n(1-\omega_n) \leq \exp\Big(-\tfrac{1}{2}m\big(\|Q_\theta\| - 3\delta\big)^2 + m\log\|Q_\theta\|^N\Big).$$

In the case that $\|Q_\theta\| < 1$, we see that:

$$Q_\theta^n(1-\omega_n) \leq e^{-\frac{1}{2}m\delta^2}$$

In the case that $\|Q_\theta\| \geq 1$, we use the identity $\log q \leq q-1$ and the fact that $-\frac{1}{2}(q-3\delta)^2+(q-1)$ has no zeroes for $q \in [1,\infty)$ if we choose $\delta < 1/6$, to conclude that the exponent is negative and bounded away from 0:

$$Q_\theta^n(1-\omega_n) \leq e^{-mc}.$$

for some $c > 0$. Combining the two last displayed bounds leads to the assertion, if we notice that $m = (n-r)/N \geq n/N - 1$, absorbing eventual constants multiplying the exponential factor in (2.23) by a slightly lower choice of D (and for large enough n). □

The following lemma is used in the proof of theorem 2.3 to control the behaviour of $\|Q_\theta\|$ in neighbourhoods of θ^*.

Lemma 2.7. *Assume that $P_0(p_\theta/p_{\theta^*})$ and $-P_0\log(p_\theta/p_0)$ are finite for all θ in a neighbourhood U' of θ^*. Furthermore, assume that there exist a measurable function m such that*

$$\Big|\log\frac{p_\theta}{p_{\theta^*}}\Big| \leq m\|\theta-\theta^*\|, \quad (P_0 - a.s.). \tag{2.34}$$

for all $\theta \in U'$ and such that $P_0(e^{sm}) < \infty$ for some $s > 0$. Then

$$P_0\Big|\frac{p_\theta}{p_{\theta^*}} - 1 - \log\frac{p_\theta}{p_{\theta^*}} - \frac{1}{2}\Big(\log\frac{p_\theta}{p_{\theta^*}}\Big)^2\Big| = o\big(\|\theta-\theta^*\|^2\big).$$

Proof The function $R(x)$ defined by $e^x = 1 + x + \frac{1}{2}x^2 + x^2\,R(x)$ increases from $-\frac{1}{2}$ in the limit $(x \to -\infty)$ to ∞ as $(x \to \infty)$, with $R(x) \to R(0) = 0$ if $(x \to 0)$. We also have $|R(-x)| \leq R(x) \leq e^x/x^2$ for all $x > 0$. The *l.h.s.* of the assertion of the lemma can be written as

$$P_0\Big(\log\frac{p_\theta}{p_{\theta^*}}\Big)^2\Big|R\Big(\log\frac{p_\theta}{p_{\theta^*}}\Big)\Big| \leq \|\theta-\theta^*\|^2 P_0\big(m^2\,R(m\|\theta-\theta^*\|)\big).$$

The expectation on the *r.h.s.* of the above display is bounded by $P_0 m_\theta^2 R(\epsilon m_\theta)$ if $\|\theta-\theta^*\| \leq \epsilon$. The functions $m^2R(\epsilon m)$ are dominated by e^{sm} for sufficiently small ϵ and converge pointwise to $m^2R(0) = 0$ as $\epsilon \downarrow 0$. The lemma then follows from the dominated convergence theorem. □

2.4 Consistency and testability

The conditions for the theorems concerning rates of convergence and limiting behaviour of the posterior distribution discussed in the previous sections include several requirements on the model involving the true distribution P_0. Depending on the specific model and true distribution, these requirements may be rather stringent, disqualifying for instance models in which $-P_0 \log p_\theta / p_{\theta^*} = \infty$ for θ in neighbourhoods of θ^*. To drop this kind of condition from the formulation and nevertheless maintain the current proof(s), we have to find other means to deal with 'undesirable' subsets of the model. In this section we show that if Kullback-Leibler neighbourhoods of the point of convergence receive enough prior mass and asymptotically consistent uniform tests for P_0 versus such subsets exist, they can be excluded from the model beforehand. As a special case, we derive a misspecified version of Schwartz' consistency theorem (see Schwartz (1965) [82]). Results presented in this section hold for the parametric models considered in previous sections, but are also valid in non-parametric situations.

2.4.1 Exclusion of testable model subsets

We start by formulating and proving the lemma announced above, in its most general form. Specializing to less general situations, we derive a corollary that can be used in most circumstances and we cast the lemma in a form reminiscent of Schwartz's consistency theorem, asserting that the posterior concentrates its mass in every neighbourhood of the point(s) at minimal Kullback-Leibler divergence with respect to the true distribution P_0.

Lemma 2.8. *Let $V \subset \Theta$ be a (measurable) subset of the model Θ. Assume that for some $\epsilon > 0$:*

$$\Pi\Big(\theta \in \Theta \,:\, -P_0 \log \frac{p_\theta}{p_{\theta^*}} \leq \epsilon\Big) > 0, \tag{2.35}$$

and there exist a constant $\beta > \epsilon$ and a sequence $(\phi_n)_{n\geq 1}$ of test-functions such that:

$$P_0^n \phi_n \to 0, \qquad \sup_{\theta \in V} Q_\theta^n (1 - \phi_n) \leq e^{-n\beta} \tag{2.36}$$

for large enough $n \geq 1$. Then:

$$P_0^n \Pi\big(V | X_1, X_2, \ldots, X_n\big) \to 0.$$

Proof We start by splitting the P_0^n-expectation of the posterior measure of V with the test function ϕ_n and taking the limes superior:

$$\begin{aligned}
&\limsup_{n\to\infty} P_0^n \Pi\big(V | X_1, X_2, \ldots, X_n\big) \\
&\qquad \leq \limsup_{n\to\infty} P_0^n \Pi\big(V | X_1, X_2, \ldots, X_n\big)(1 - \phi_n) + \limsup_{n\to\infty} P_0^n \phi_n \\
&\qquad = \limsup_{n\to\infty} P_0^\infty \Pi\big(V | X_1, X_2, \ldots, X_n\big)(1 - \phi_n) \\
&\qquad \leq P_0^\infty \Big(\limsup_{n\to\infty} \Pi\big(V | X_1, X_2, \ldots, X_n\big)(1 - \phi_n) \Big)
\end{aligned} \tag{2.37}$$

by Fatou's lemma. We therefore concentrate on the quantities

$$\begin{aligned}\limsup_{n\to\infty} &\Pi_n(V|X_1,X_2,\ldots,X_n)(1-\phi_n)(X_1,X_2,\ldots,X_n)\\ &= \limsup_{n\to\infty} \frac{\displaystyle\int_V \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,(1-\phi_n)(X_1,X_2,\ldots,X_n)\,d\Pi(\theta)}{\displaystyle\int_\Theta \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,d\Pi(\theta)},\end{aligned} \tag{2.38}$$

and, more particularly, show that suitable upper and lower bounds exist for the numerator and denominator of the fraction on the *r.h.s.* of the last display in the limit that $n \to \infty$.

Starting with the denominator, we consider the subset $K_\epsilon = \{\theta \in \Theta : -P_0 \log(p_\theta/p_{\theta^*}) \leq \epsilon\}$. For every $\theta \in K_\epsilon$, the strong law of large numbers says that:

$$\left| \mathbb{P}_n \log \frac{p_\theta}{p_{\theta^*}} - P_0 \log \frac{p_\theta}{p_{\theta^*}} \right| \to 0, \quad (P_0 - a.s.).$$

Hence for every $\alpha > \epsilon$ and all $P \in K_\epsilon$, there exists an $N \geq 1$ such that for all $n \geq N$, $\prod_{i=1}^n (p_\theta/p_{\theta^*})(X_i) \geq e^{-n\alpha}$, P_0^n-almost-surely. This can be used to lower-bound the denominator of (2.38) P_0^n-almost-surely as follows:

$$\begin{aligned}\liminf_{n\to\infty} e^{n\alpha} \int_\Theta \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,d\Pi(\theta) &\geq \liminf_{n\to\infty} e^{n\alpha} \int_{K_\epsilon} \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,d\Pi(\theta)\\ &\geq \int_{K_\epsilon} \liminf_{n\to\infty} e^{n\alpha} \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,d\Pi(\theta) \geq \Pi(K_\epsilon),\end{aligned}$$

where we use Fatou's lemma to obtain the second inequality. Since by assumption, $\Pi(K_\epsilon) > 0$ we see that:

$$\begin{aligned}\limsup_{n\to\infty} &\frac{\displaystyle\int_V \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,(1-\phi_n)(X_1,X_2,\ldots,X_n)\,d\Pi(\theta)}{\displaystyle\int_\Theta \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,d\Pi(\theta)}\\ &\leq \frac{\limsup_{n\to\infty} e^{n\alpha} \displaystyle\int_V \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,(1-\phi_n)(X_1,X_2,\ldots,X_n)\,d\Pi(\theta)}{\liminf_{n\to\infty} e^{n\alpha} \displaystyle\int_\Theta \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,d\Pi(\theta)}\\ &\leq \frac{1}{\Pi(K_\epsilon)} \limsup_{n\to\infty} f_n(X_1,X_2,\ldots,X_n),\end{aligned} \tag{2.39}$$

where we use the following, P_0^∞-almost-surely defined sequence of random variables $(f_n)_{n\geq 1}$, $f_n : \mathscr{X}^n \to \mathbb{R}$:

$$f_n(X_1,X_2,\ldots,X_n) = e^{n\alpha} \int_V \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,(1-\phi_n)(X_1,X_2,\ldots,X_n)\,d\Pi(\theta).$$

Fubini's theorem and the assumption that the test-sequence is uniformly exponential, guarantee that for large enough n,

$$P_0^\infty f_n = P_0^n f_n = e^{n\alpha} \int_V Q_\theta^n (1 - \phi_n)\, d\Pi(\theta) \leq e^{-n(\beta-\alpha)}.$$

Markov's inequality can then be used to show that:

$$P_0^\infty \big(f_n > e^{-\frac{n}{2}(\beta-\epsilon)} \big) \leq e^{n(\alpha - \frac{1}{2}(\beta+\epsilon))}.$$

Since $\beta > \epsilon$, we can choose α such that $\epsilon < \alpha < \frac{1}{2}(\beta + \epsilon)$ so that the series $\sum_{n=1}^\infty P_0^\infty(f_n > \exp -\frac{n}{2}(\beta - \epsilon))$ converges. The first Borel-Cantelli lemma then leads to the conclusion that:

$$P_0^\infty \Big(\bigcap_{N=1}^\infty \bigcup_{n \geq N} \{ f_n > e^{-\frac{n}{2}(\beta-\epsilon)} \} \Big) = P_0^\infty \Big(\limsup_{n\to\infty} \big(f_n - e^{-\frac{n}{2}(\beta-\epsilon)} \big) > 0 \Big) = 0$$

Since $f_n \geq 0$, we see that $f_n \to 0$, $(P_0 - a.s.)$, which we substitute in (2.39) and subsequently in (2.38) and (2.37) to conclude the proof. □

In many situations, (2.35) is satisfied for every $\epsilon > 0$. In that case the construction of uniform exponentially powerful tests from asymptotically consistent tests (as demonstrated in the proof of lemma 2.6) can be used to fulfil (2.36) under the condition that an asymptotically consistent uniform test-sequence exists.

Corollary 2.1. *Let $V \subset \Theta$ be a (measurable) subset of the model Θ. Assume that for all $\epsilon > 0$ (2.35) is satisfied and that there exists a test-sequence $(\phi_n)_{n\geq 1}$ such that:*

$$P_0^n \phi_n \to 0, \qquad \sup_{\theta \in V} Q_\theta^n (1 - \phi_n) \to 0 \tag{2.40}$$

Then

$$P_0^n \Pi\big(V \;\big|\; X_1, X_2, \ldots, X_n \big) \to 0.$$

In this corollary form, the usefulness of lemma 2.8 is most apparent. All subsets V of the model that can be distinguished from P_0 based on a characteristic property (formalised by the test functions above) in a uniform manner (*c.f.* (2.40)) may be discarded from proofs like that of theorem 2.2: before the first step in the proof, we split up the posterior by intersecting with V and with its complement. The assertion of the above corollary guarantees that the first term converges to zero, leaving us to give the proof only for θ in the complement of V. Hence the properties *assumed* in the statement of (for instance) theorem 2.2, can be left out as conditions *if* a suitable test sequence exist.

Whether or not a suitable test sequence can be found depends on the particular model and true distribution in question and little can be said in any generality. We discuss one construction in particular though: the following lemma demonstrates that a subset of the model in which the Kullback-Leibler divergence differs from the minimal value in the model sufficiently in a uniform manner, can be excluded on the basis of lemma 2.8.

Lemma 2.9. *Let $V \in \Theta$ be a (measurable) subset of the model Θ. Assume that for all $\epsilon > 0$ (2.35) is satisfied and suppose that there exists a sequence $(M_n)_{n\geq 1}$ of positive numbers such that $M_n \to \infty$ and*

$$P_0^n\Big(\inf_{\theta\in V} -\mathbb{P}_n \log \frac{p_\theta}{p_{\theta^*}} < \frac{1}{n} M_n \Big) \to 0. \tag{2.41}$$

Then

$$P_0^n \Pi\big(V \;\big|\; X_1, X_2, \ldots, X_n \big) \to 0.$$

Proof Define the sequence of test functions:

$$\psi_n = 1\Big\{ \inf_{\theta\in V} -\mathbb{P}_n \log \frac{p_\theta}{p_{\theta^*}} < \frac{1}{n} M_n \Big\}$$

According to assumption (2.41), $P_0^n \psi_n \to 0$. Let $\theta \in V$ be given.

$$\begin{aligned} Q_\theta^n(1-\psi_n) &= P_0^n\Big(\frac{dP_\theta^n}{dP_{\theta^*}^n}(1-\psi) \Big) = P_0^n\Big(\frac{dP_\theta^n}{dP_{\theta^*}^n} 1\Big\{ \sup_{\theta\in V} n\mathbb{P}_n \log \frac{p_\theta}{p_{\theta^*}} \leq -M_n \Big\} \Big) \\ &= P_0^n\Big(\frac{dP_\theta^n}{dP_{\theta^*}^n} 1\Big\{ \sup_{\theta\in V} \log \frac{dP_\theta^n}{dP_{\theta^*}^n} \leq -M_n \Big\} \Big) \leq e^{-M_n} P_0^n\Big(\sup_{\theta\in V} \log \frac{dP_\theta^n}{dP_{\theta^*}^n} \leq -M_n \Big) \\ &\leq e^{-M_n} \to 0. \end{aligned}$$

Since M_n does not depend on θ, convergence to 0 is uniform over V. Corollary 2.1 then gives the assertion. □

Finally, we note that lemma 2.8 can also be used to prove consistency. Presently, we do not assume that there exists a *unique* point of minimal Kullback-Leibler divergence; we define Θ^* to be the set of points in the model at minimal Kullback-Leibler divergence with respect to the true distribution P_0, $\Theta^* = \{\theta \in \Theta : -P_0 \log(p_\theta/p_0) = \inf_\Theta -P_0 \log(p_\theta/p_0)\}$ (assuming, of course, that this set is measurable), and we consider the posterior probability of this set under the conditions of corollary 2.1. We write $d(\theta, \Theta^*)$ for the infimum of $\|\theta - \theta^*\|$ over $\theta^* \in \Theta^*$.

Corollary 2.2. *(Schwartz consistency) Assume that for all $\epsilon > 0$ (2.35) is satisfied and that for all $\eta > 0$ there exists a test-sequence $(\phi_n)_{n\geq 1}$ such that:*

$$P_0^n \phi_n \to 0, \qquad \sup_{\theta : d(\theta,\Theta^*) > \eta} Q_\theta^n(1-\phi_n) \to 0$$

Then

$$P_0^n \Pi\big(\Theta^* \;\big|\; X_1, X_2, \ldots, X_n \big) \to 1.$$

2.5 Three lemmas used in the main proof

In this section, we state and prove three lemmas that are used in the proof of theorem 2.1. The first lemma shows that for a sequence of continuous random functions, uniform convergence on a compact set is equivalent to convergence for all sequences in the compactum. The other two lemmas serve to extend convergence in total variation for a growing sequence of compacta as obtained in (2.10) to the assertion of the theorem.

Lemma 2.10. *Let $(f_n)_{n\geq 1}$ be a sequence of random functions $K \to \mathbb{R}$, where K is compact. Assume that for large enough $n \geq 1$, f_n is continuous P_0^n-almost-surely. Then the following are equivalent*[3]*:*

(i) Uniform convergence in probability:

$$\sup_{h\in K}\big|f_n(h)\big| \xrightarrow{P_0} 0, \quad (n \to \infty),$$

(ii) For any random sequence $(h_n)_{n\geq 1} \subset K$:

$$f_n(h_n) \xrightarrow{P_0} 0, \quad (n \to \infty),$$

Proof *((ii)⇒(i), by contradiction.)* Assume that there exist $\delta, \epsilon > 0$ such that:

$$\limsup_{n\to\infty} P_0\big(\sup_{h\in K}\big|f_n(h)\big| > \delta\big) = \epsilon.$$

Since the functions f_n are continuous P_0-almost-surely, there exists (with P_0-probability one) a sequence $(\tilde{h}_n)_{n\geq 1}$ such that for every $n \geq 1$, $\tilde{h}_n \in K$ and

$$\big|f_n(\tilde{h}_n)\big| = \sup_{h\in K}\big|f_n(h)\big|.$$

Consequently, for this particular random sequence in K, we have:

$$\limsup_{n\to\infty} P_0\Big(\big|f_n(\tilde{h}_n)\big| > \delta\Big) = \epsilon > 0.$$

which contradicts (ii). *((i)⇒(ii).)* Given a random sequence $(h_n)_{n\geq 1} \subset K$, and for every $\delta > 0$,

$$P_0\Big(\sup_{h\in K}\big|f_n(h)\big| > \delta\Big) \geq P_0\Big(\big|f_n(h_n)\big| > \delta\Big).$$

Given (i), the *l.h.s.* converges to zero and hence so does the *r.h.s.*. □

The next lemma shows that given two sequences of probability measures, a sequence of balls that grows fast enough can be used conditionally to calculate the difference in total-variational distance, even when the sequences consist of random measures.

Lemma 2.11. *Let $(\Pi_n)_{n\geq 1}$ and $(\Phi_n)_{n\geq 1}$ be two sequences of random probability measures on $\mathbb{R}^d$. Let $(K_n)_{n\geq 1}$ be a sequence of subsets of $\mathbb{R}^d$ such that*

$$\Pi_n(\mathbb{R}^d \setminus K_n) \xrightarrow{P_0} 0, \quad \Phi_n(\mathbb{R}^d \setminus K_n) \xrightarrow{P_0} 0. \tag{2.42}$$

Then

$$\big\|\Pi_n - \Phi_n\big\| - \big\|\Pi_n^{K_n} - \Phi_n^{K_n}\big\| \xrightarrow{P_0} 0. \tag{2.43}$$

[3]Measurability of the (possibly uncountable) supremum is guaranteed if K is a subset of a metric space; in that case totally boundedness assures the existance of a dense countable subset K' so that:

$$\sup_{h\in K} F(h) = \sup_{h'\in K'} F(h'),$$

for every continuous function $F : K \to \mathbb{R}$. Measurability of the $\tilde{h}_n$, $(n \geq 1)$ used in the proof is more difficult to establish.

Proof Let K, a measurable subset of $\mathbb{R}^d$ and $n \geq 1$ be given and assume that $\Pi_n(K) > 0$ and $\Phi_n(K) > 0$. Then for any measurable $B \subset \mathbb{R}^d$ we have:

$$\begin{aligned}
\big|\Pi_n(B) - \Pi_n^K(B)\big| &= \Big|\Pi_n(B) - \frac{\Pi_n(B \cap K)}{\Pi_n(K)}\Big| \\
&= \big|\Pi_n\big(B \cap (\mathbb{R}^d \setminus K)\big) + \big(1 - \Pi_n(K)^{-1}\big)\Pi_n(B \cap K)\big| \\
&= \big|\Pi_n\big(B \cap (\mathbb{R}^d \setminus K)\big) - \Pi_n(\mathbb{R}^d \setminus K)\Pi_n^K(B)\big| \\
&\leq \Pi_n\big(B \cap (\mathbb{R}^d \setminus K)\big) + \Pi_n(\mathbb{R}^d \setminus K)\Pi_n^K(B) \\
&\leq 2\Pi_n(\mathbb{R}^d \setminus K).
\end{aligned}$$

and hence also:

$$\Big|\big(\Pi_n(B) - \Pi_n^K(B)\big) - \big(\Phi_n(B) - \Phi_n^K(B)\big)\Big| \leq 2\big(\Pi_n(\mathbb{R}^d \setminus K) + \Phi_n(\mathbb{R}^d \setminus K)\big). \tag{2.44}$$

As a result of the triangle inequality, we then find that the difference in total-variation distances between Π_n and Φ_n on the one hand and Π_n^K and Φ_n^K on the other is bounded above by the expression on the right in the above display (which is independent of B).

Define A_n, B_n to be the events that $\Pi_n(K_n) > 0$, $\Phi_n(K_n) > 0$ respectively. On $\Xi_n = A_n \cap B_n$, $\Pi_n^{K_n}$ and $\Phi_n^{K_n}$ are well-defined probability measures. Assumption (2.42) guarantees that $P_0^n(\Xi_n)$ converges to 1. Restricting attention to the event Ξ_n in the above upon substitution of the sequence $(K_n)_{n\geq 1}$ and using (2.42) for the limit of (2.44) we find (2.43), where it is understood that the conditional probabilities on the *l.h.s.* are well-defined with probability growing to 1. □

To apply the above lemma in the concluding steps of the proof of theorem 2.1, rate conditions for both posterior and limiting normal sequences are needed. The rate condition (2.6) for the posterior is assumed and the following lemma demonstrates that its analog for the sequence of normals is satisfied when the sequence of centre points Δ_{n,θ^*} is uniformly tight.

Lemma 2.12. *Let K_n be a sequence of balls centred on the origin with radii $M_n \to \infty$. Let $(\Phi_n)_{n\geq 1}$ be a sequence of normal distributions (with fixed covariance matrix V) located respectively at the (random) points $(\Delta_n)_{n\geq 1} \subset \mathbb{R}^d$. If the sequence Δ_n is uniformly tight, then:*

$$\Phi_n(\mathbb{R}^d \setminus K_n) = N_{\Delta_n, V}(\mathbb{R}^d \setminus K_n) \xrightarrow{P_0} 0.$$

proof Let $\delta > 0$ be given. Uniform tightness of the sequence $(\Delta_n)_{n\geq 1}$ implies the existence of a constant $L > 0$ such that:

$$\sup_{n\geq 1} P_0^n(\|\Delta_n\| \geq L) \leq \delta.$$

For all $n \geq 1$, call $A_n = \{\|\Delta_n\| \geq L\}$. Let $\mu \in \mathbb{R}^d$ be given. Since $N(\mu, V)$ is tight, there exists for every given $\epsilon > 0$ a constant L' such that $N_{\mu,V}(B(\mu, L')) \geq 1 - \epsilon$ (where $B(\mu, L')$

defines a ball of radius L' around the point μ. Assuming that $\mu \leq L$, $B(\mu, L') \subset B(0, L + L')$ so that with $M = L + L'$, $N_{\mu,V}(B(0, M)) \geq 1 - \epsilon$ for all μ such that $\|\mu\| \leq L$. Choose $N \geq 1$ such that $M_n \geq M$ for all $n \geq N$. Let $n \geq N$ be given. Then:

$$\begin{aligned} P_0^n\big(\Phi_n(\mathbb{R}^d \setminus B(0, M_n) > \epsilon\big) &\leq P_0^n\big(A_n\big) + P_0^n\Big(\big\{\Phi_n(\mathbb{R}^d \setminus B(0, M_n) > \epsilon\big\} \cap A_n^c\Big) \\ &\leq \delta + P_0^n\Big(\big\{N_{\Delta_n,V}(B(0, M_n)^c) > \epsilon\big\} \cap A_n^c\Big) \end{aligned} \tag{2.45}$$

Note that on the complement of A_n, $\|\Delta_n\| < L$, so:

$$N_{\Delta_n,V}(B(0, M_n)^c) \leq 1 - N_{\Delta_n,V}(B(0, M)) \leq 1 - \inf_{\|\mu\| \leq L} N_{\mu,V}(B(0, M)) \leq \epsilon,$$

and we conclude that the last term on the *r.h.s.* of (2.45) equals zero. □

Chapter 3

Misspecification in non-parametric Bayesian statistics

Often, estimation procedures behave far more erratic when used in a non-parametric than in a parametric model. For instance, where maximum-likelihood estimation is a straightforward and generally well-defined procedure in parametric situations, its behaviour is far more difficult to control in non-parametric models. As far as rates of convergence are concerned, we recall the relation that exists between metric entropy numbers and optimality of rates. In non-parametric models, restrictions like (1.39) play a far-more-serious role.

The relative difficulty of infinite-dimensional estimation has its consequences for Bayesian methods as well. To begin with, the definition of prior measures on infinite-dimensional collections of probability measures is non-trivial (see *e.g.* Ghosh and Ramamoorthi (2003) [41]). On the other hand, Schwartz' consistency theorem (theorem 1.7) and the rates-of-convergence theorem 1.8 are applicable in non-parametric models. Both rely on conditions involving the prior mass of certain Kullback-Leibler neighbourhoods of P_0 and the existence of suitable test sequences. As pointed out in subsection 1.3.2, condition (1.39) for the optimal (Hellinger) rate can be related to an existence proof for suitable tests.

In this chapter, we formulate a number of theorems (*e.g.* theorems 3.1 and 3.2–3.4) concerning Bayesian rates of convergence under misspecification. Although a number of new difficulties arise (most importantly, the somewhat complicated relation between so-called *entropy numbers for testing under misspecification* (see the definition preceding inequality (3.4)) and ordinary metric entropy numbers; see lemma 3.1) the basic structure of the conditions remains as described above, including the relation between entropy numbers and the existence of suitable test sequences. In many cases, the rate of convergence under misspecification is the *same* rate that is achieved in the case of a well-specified model. The results are applied to a number of models, including Gaussian mixtures and non-parametric regression. The presentation of these results as found in the remainder of this chapter has been (tentatively) accepted for publication in the *Annals of Statistics*.

Misspecification in non-parametric Bayesian statistics

B.J.K. KLEIJN AND A.W. VAN DER VAART

Free University Amsterdam

Abstract

We consider the asymptotic behavior of posterior distributions if the model is misspecified. Given a prior distribution and a random sample from a distribution P_0, which may not be in the support of the prior, we show that the posterior concentrates its mass near the points in the support of the prior that minimize the Kullback-Leibler divergence with respect to P_0. An entropy condition and a prior-mass condition determine the rate of convergence. The method is applied to several examples, with special interest for infinite-dimensional models. These include Gaussian mixtures, nonparametric regression and parametric models.

3.1 Introduction

Of all criteria for statistical estimation, asymptotic consistency is among the least disputed. Consistency requires that the estimation procedure comes arbitrarily close to the true, underlying distribution, if enough observations are used. It is of a frequentist nature, because it presumes a notion of an underlying, true distribution for the observations. If applied to posterior distributions it is also considered a useful property by many Bayesians, as it could warn one away from prior distributions with undesirable, or unexpected, consequences. Priors which lead to undesirable posteriors have been documented in particular for non- or semi-parametric models (see *e.g.* Diaconis and Freedman (1986) [24, 25]), in which case it is also difficult to motivate a particular prior on purely intuitive, subjective grounds.

In the present paper we consider the situation that the posterior distribution cannot possibly be asymptotically consistent, because the model, or the prior, is misspecified. From a frequentist point of view the relevance of studying misspecification is clear, because the

assumption that the model contains the true, underlying distribution may lack realistic motivation in many practical situations. From an objective Bayesian point of view the question is of interest, because in principle the Bayesian paradigm allows unrestricted choice of a prior, and hence we must allow for the possibility that the fixed distribution of the observations does not belong to the support of the prior. In this paper we show that in such a case the posterior will concentrate near a point in the support of the prior that is closest to the true sampling distribution as measured through the Kullback-Leibler divergence, and we give a characterization for the rate of concentration near this point.

Throughout the paper we assume that $X_1, X_2, \ldots$ are i.i.d. observations, each distributed according to a probability measure P_0. Given a model $\mathscr{P}$ and a prior Π, supported on $\mathscr{P}$, the posterior mass of a measurable subset $B \subset \mathscr{P}$ is given by:

$$\Pi_n\big(B \,\big|\, X_1, \ldots, X_n \big) = \int_B \prod_{i=1}^n p(X_i)\, d\Pi(P) \,\Big/\, \int_{\mathscr{P}} \prod_{i=1}^n p(X_i)\, d\Pi(P). \tag{3.1}$$

Here it is assumed that the model is dominated by a σ-finite measure μ, and the density of a typical element $P \in \mathscr{P}$ relative to the dominating measure is written p and assumed appropriately measurable. If we assume that the model is well specified, *i.e.* $P_0 \in \mathscr{P}$, then posterior consistency means that the posterior distributions concentrate an arbitrarily large fraction of their total mass in arbitrarily small neighborhoods of P_0, if the number of observations used to determine the posterior is large enough. To formalise this, we let d be a metric on $\mathscr{P}$ and say that the Bayesian procedure for the specified prior is *consistent*, if for every $\epsilon > 0$, $\Pi_n(\{P : d(P, P_0) > \epsilon\}|X_1, \ldots, X_n) \to 0$, in P_0-probability. More specific information concerning the asymptotic behaviour of an estimator is given by its rate of convergence. Let $\epsilon_n > 0$ be a sequence that decreases to zero and suppose that there exists a constant $M > 0$ such that:

$$\Pi_n\big(P \in \mathscr{P} : d(P, P_0) > M\epsilon_n \,\big|\, X_1, \ldots, X_n \big) \to 0, \tag{3.2}$$

in P_0-probability. The sequence ϵ_n corresponds to a decreasing sequence of neighborhoods of P_0, the d-radius of which goes to zero with n, while still capturing most of the posterior mass. If (3.2) is satisfied, then we say that the rate of convergence is at least ϵ_n.

If P_0 is at a positive distance from the model $\mathscr{P}$ and the prior concentrates all its mass on $\mathscr{P}$, then the posterior is inconsistent as it will concentrate all its mass on $\mathscr{P}$ as well. However, in this paper we show that the posterior will still settle down near a given measure $P^* \in \mathscr{P}$, and we shall characterize the sequences ϵ_n such that the preceding display is valid with $d(P, P^*)$ taking the place of $d(P, P_0)$.

One would expect the posterior to concentrate its mass near minimum Kullback-Leibler points, since asymptotically the likelihood $\prod_{i=1}^n p(X_i)$ is maximal near points of minimal Kullback-Leibler divergence. The integrand in the numerator of (3.1) is the likelihood, so subsets of the model in which the (log-)likelihood is large account for a large fraction of the total posterior mass. Hence it is no great surprise that the appropriate point of convergence P^* is a minimum Kullback-Leibler point in $\mathscr{P}$, but the general issue of rates (and which

metric d to use) turns out to be more complicated than expected. We follow the work by Ghosal *et al.* (2000) [39] for the well-specified situation, but need to adapt, change or extend many steps.

After deriving general results, we consider several examples in some detail, including Bayesian fitting of Gaussian mixtures using Dirichlet priors on the mixing distribution, the regression problem, and parametric models. Our results on the regression problem allows one, for instance, to conclude that a Bayesian approach in the nonparametric problem that uses a prior on the regression function, but employs a normal distribution for the errors, will lead to consistent estimation of the regression function, even if the regression errors are non-Gaussian. This result, which is the Bayesian counterpart of the well-known fact that least squares estimators (the maximum likelihood estimators if the errors are Gaussian) perform well even if the errors are non-Gaussian, is important to validate the Bayesian approach to regression but appears to have received little attention in the literature.

A few notes concerning notation and organization are in order. Let $L_1(\mathscr{X}, \mathscr{A})$ denote the set of all finite signed measures on $(\mathscr{X}, \mathscr{A})$ and let $\mathrm{co}(\mathscr{Q})$ be the convex hull of a set of measures $\mathscr{Q}$: the set of all finite linear combinations $\sum_i \lambda_i Q_i$ for $Q_i \in \mathscr{Q}$ and $\lambda_i \geq 0$ with $\sum_i \lambda_i = 1$. For a measurable function f let Qf denote the integral $\int f\, dQ$. The paper is organized as follows. Section 3.2 contains the main results of the paper, in increasing generality. Sections 3.3, 3.4, and 3.5 concern the three classes of examples that we consider: mixtures, the regression model, and parametric models. Sections 3.6 and 3.7 contain the proofs of the main results, where the necessary results on tests are developed in section 3.6 and are of independent interest. The final section is a technical appendix.

3.2 Main results

Let $X_1, X_2, \ldots$ be an *i.i.d.* sample from a distribution P_0 on a measurable space $(\mathscr{X}, \mathscr{A})$. Given a collection $\mathscr{P}$ of probability distributions on $(\mathscr{X}, \mathscr{A})$ and a prior probability measure Π on $\mathscr{P}$, the posterior measure is defined as in (3.1) (where $0/0 = 0$ by definition). Here it is assumed that the 'model' $\mathscr{P}$ is dominated by a σ-finite measure μ and that $x \mapsto p(x)$ is a density of $P \in \mathscr{P}$ relative to μ such that the map $(x, p) \mapsto p(x)$ is measurable, relative to the product of $\mathscr{A}$ and an appropriate σ-field on $\mathscr{P}$, so that the right side of (3.1) is a measurable function of $(X_1, \ldots, X_n)$ and a probability measure as a function of B for every $X_1, \ldots, X_n$ such that the denominator is positive. The 'true' distribution P_0 may or may not belong to the model $\mathscr{P}$. For simplicity of notation we assume that P_0 possesses a density p_0 relative to μ as well. (Alternatively, fractions of densities may be replaced by Radon-Nikodym derivatives throughout, eliminating the need for a dominating measure.)

Informally we think of the model $\mathscr{P}$ as the 'support' of the prior Π, but we shall not make this precise in a topological sense. At this point we only assume that the prior concentrates on $\mathscr{P}$ in the sense that $\Pi(\mathscr{P}) = 1$ (but we note later that this too can be relaxed). Further requirements are made in the statements of the main results. Our main theorems implicitly

assume the existence of a point $P^* \in \mathscr{P}$ minimizing the Kullback-Leibler divergence of P_0 to the model $\mathscr{P}$. In particular, the minimal Kullback-Leibler divergence is assumed to be finite, *i.e.* P^* satisfies:

$$-P_0 \log \frac{p^*}{p_0} < \infty. \tag{3.3}$$

By the convention that $\log 0 = -\infty$, the above implies that $P_0 \ll P^*$ and hence we assume without loss of generality that the density p^* is strictly positive at the observations.

Our theorems give sufficient conditions for the posterior distribution to concentrate in neighborhoods of P^* at a rate that is determined by the amount of prior mass 'close to' the minimal Kullback-Leibler point P^* and the 'entropy' of the model. To specify the terms between quotations marks, we make the following definitions.

We define the entropy and the neighborhoods in which the posterior is to concentrate its mass relative to a semi-metric d on $\mathscr{P}$. The general results are formulated relative to an arbitrary semi-metric and next the conditions will be simplified for more specific choices. Whether or not these simplifications can be made depends on the model $\mathscr{P}$, convexity being an important special case (see lemma 3.2). Unlike the case of well-specified priors, considered *e.g.* in Ghosal *et al.* (2000) [39], the Hellinger distance is not always appropriate in the misspecified situation. The general entropy bound is formulated in terms of a *covering number for testing under misspecification*, defined for $\epsilon > 0$ as follows: we define $N_t(\epsilon, \mathscr{P}, d; P_0, P^*)$ as the minimal number N of convex sets $B_1, \ldots, B_N$ of probability measures on $(\mathscr{X}, \mathscr{A})$ needed to cover the set $\{P \in \mathscr{P} : \epsilon < d(P, P^*) < 2\epsilon\}$ such that, for every i,

$$\inf_{P \in B_i} \sup_{0<\alpha<1} -\log P_0\Big(\frac{p}{p^*}\Big)^\alpha \geq \frac{\epsilon^2}{4}. \tag{3.4}$$

If there is no finite covering of this type we define the covering number to be infinite. We refer to the logarithms $\log N_t(\epsilon, \mathscr{P}, d; P_0, P^*)$ as *entropy numbers for testing under misspecification*. Because the measures P_0 and P^* are fixed in the following, we may delete them from the notation and write the covering number as $N_t(\epsilon, \mathscr{P}, d)$. These numbers differ from ordinary metric entropy numbers in that the covering sets B_i are required to satisfy the preceding display rather than to be balls of radius ϵ. We insist that the sets B_i be convex and that (3.4) hold for every $P \in B_i$. This implies that (3.4) may involve measures P that do not belong to the model $\mathscr{P}$ if this is not convex itself.

For $\epsilon > 0$ we define a specific kind of Kullback-Leibler neighborhoods of P^* by

$$B(\epsilon, P^*; P_0) = \Big\{P \in \mathscr{P} : -P_0 \log \frac{p}{p^*} \leq \epsilon^2,\ P_0\Big(\log \frac{p}{p^*}\Big)^2 \leq \epsilon^2\Big\}. \tag{3.5}$$

Theorem 3.1. *For a given model $\mathscr{P}$ with prior Π and some $P^* \in \mathscr{P}$ assume that the Kullback-Leibler divergence with respect to P_0 $-P_0 \log(p^*/p_0)$ is finite and that $P_0(p/p^*) < \infty$ for all $P \in \mathscr{P}$. Suppose that there exists a sequence of strictly positive numbers ϵ_n with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ and a constant $L > 0$, such that for all n:*

$$\Pi\big(B(\epsilon_n, P^*; P_0)\big) \geq e^{-Ln\epsilon_n^2}, \tag{3.6}$$

$$N_t\big(\epsilon, \mathscr{P}, d; P_0, P^*\big) \leq e^{n\epsilon_n^2}, \qquad \text{for all } \epsilon > \epsilon_n. \tag{3.7}$$

Then for every sufficiently large constant M, as $n \to \infty$,

$$\Pi_n\big(P \in \mathscr{P} : d(P, P^*) \geq M\epsilon_n \;\big|\; X_1, \ldots, X_n \big) \to 0, \qquad \text{in } L_1(P_0^n). \tag{3.8}$$

The proof of this theorem is given in section 3.7. The theorem does not explicitly require that P^* is a point of minimal Kullback-Leibler divergence, but this is implied by the conditions (see lemma 3.16 below). The theorem is extended to the case of non-unique minimal Kullback-Leibler points in section 3.2.4.

The two main conditions of theorem 3.1 are a prior mass condition (3.6) and an entropy condition (3.7), which can be compared to Schwartz' conditions for posterior consistency (see Schwartz (1965) [82]), or the two main conditions for the well-specified situation in Ghosal *et al.* (2000) [39]. Below we discuss the background of these conditions in turn.

The prior mass condition reduces to the corresponding condition for the correctly specified case in Ghosal *et al.* (2000) [39] if $P^* = P_0$. Because $-P_0 \log(p^*/p_0) < \infty$, we may rewrite the first inequality in the definition (3.6) of the set $B(\epsilon, P^*; P_0)$ as:

$$-P_0 \log \frac{p}{p_0} \leq -P_0 \log \frac{p^*}{p_0} + \epsilon^2.$$

Therefore the set $B(\epsilon, P^*; P_0)$ contains only $P \in \mathscr{P}$ that are within ϵ^2 of the minimal Kullback-Leibler divergence with respect to P_0 over the model. The lower bound (3.6) on the prior mass of $B(\epsilon, P^*; P_0)$ requires that the prior measure assigns a certain minimal share of its total mass to Kullback-Leibler neighborhoods of P^*. As argued in Ghosal *et al.* (2000) [39] a rough understanding of the exact form of (3.6) for the 'optimal' rate ϵ_n is that an optimal prior spreads its mass 'uniformly' over $\mathscr{P}$. In the proof of theorem 3.1, the prior mass condition serves to lower-bound the denominator in the expression for the posterior.

The background of the entropy condition (3.6) is more involved, but can be compared to a corresponding condition in the well-specified situation given in theorem 2.1 of Ghosal *et al.* (2000) [39]. The purpose of the entropy condition is to measure the complexity of the model, a larger entropy leading to a smaller rate of convergence. The entropy used in Ghosal *et al.* (2000) [39] is either the ordinary metric entropy $\log N(\epsilon, \mathscr{P}, d)$, or the local entropy $\log N\big(\epsilon/2, \{P \in \mathscr{P} : \epsilon < d(P, P_0) < 2\epsilon\}, d\big)$. For d the Hellinger distance the minimal ϵ_n satisfying $\log N(\epsilon_n, \mathscr{P}, d) = n\epsilon_n^2$ is roughly the fastest rate of convergence for estimating a density in the model $\mathscr{P}$ relative to d obtainable by any method of estimation (*c.f.* Birgé (1983) [15]). We are not aware of a concept of 'optimal rate of convergence' if the model is misspecified, but a rough interpretation of (3.7) given (3.6) would be that in the misspecified situation the posterior concentrates near the closest Kullback-Leibler point at the optimal rate pertaining to the model $\mathscr{P}$.

Misspecification requires that the complexity of the model is measured in an different, somewhat complicated way. In examples, depending on the semi-metric d, the covering numbers $N_t(\epsilon, \mathscr{P}, d; P_0, P^*)$ can be related to ordinary metric covering numbers $N(\epsilon, \mathscr{P}, d)$. For instance, we show below (see lemmas 3.1–3.3) that if the model $\mathscr{P}$ is convex, then the numbers

$N_t(\epsilon, \mathscr{P}, d; P_0, P^*)$ are bounded by the covering numbers $N(\epsilon, \mathscr{P}, d)$ if the distance $d(P_1, P_2)$ equals the Hellinger distance between the measures Q_i defined by $dQ_i = (p_0/p^*)\,dP_i$, *i.e.* a weighted Hellinger distance between P_1 and P_2. In the well-specified situation we have $P^* = P_0$, thus reducing d to the Hellinger distance (without weight factor). In that case there appears to be no use for the covering numbers as defined by (3.4) but in the general, misspecified situation they are essential, even for standard parametric models, such as the one-dimensional normal location model.

At a more technical level the entropy condition of Ghosal *et al.* (2000) [39] ensures the existence of certain tests of the measures P versus the true measure P_0. In the misspecified case it is necessary to compare the measures P to the minimal Kullback-Leibler point P^*, rather than to P_0. It turns out that the appropriate comparison is not a test of the measures P versus the measure P^* in the ordinary sense of testing, but to test the measures $Q(P)$ defined by $dQ(P) = (p_0/p^*)\,dP$ versus the measure P_0 (see (3.47)). With $\mathscr{Q}$ the set of measures $Q(P)$ where P ranges over $\mathscr{P}$ this leads to consideration of minimax testing risks of the type

$$\inf_{\phi} \sup_{Q \in \mathscr{Q}} \big(P_0^n \phi + Q^n(1-\phi)\big),$$

where the infimum is taken over all measurable functions ϕ taking values in $[0,1]$. A difference with the usual results on minimax testing risks is that the measures Q may not be probability measures (and may in fact be infinite in general).

Extending arguments of Le Cam and Birgé, we show in section 3.6 that for a convex set $\mathscr{Q}$ the minimax testing risk in the preceding display is bounded above by

$$\inf_{0<\alpha<1} \sup_{Q \in \mathscr{Q}} \rho_\alpha(P_0, Q)^n, \tag{3.9}$$

where the function $\alpha \mapsto \rho_\alpha(P_0, Q)$ is the Hellinger transform $\rho_\alpha(P, Q) = \int p^\alpha q^{1-\alpha}\,d\mu$. For $Q = Q(P)$, the Hellinger transform reduces to the map

$$\alpha \mapsto \rho_{1-\alpha}\big(Q(P), P_0\big) = P_0(p/p^*)^\alpha,$$

also encountered in (3.4). If the inequality in (3.4) is satisfied, then $P_0(p/p^*)^\alpha \le e^{-\epsilon^2/4}$ and hence the set of measures $Q(P)$ with P ranging over B_i can be tested with error probabilities bounded by $e^{-n\epsilon^2/4}$. For ϵ bounded away from zero, or converging slowly to zero, these probabilities are exponentially small, ensuring that the posterior does not concentrate on the 'unlikely alternatives' B_i.

The testing bound (3.9) is valid for convex alternatives $\mathscr{Q}$, but the alternatives of interest $\{P \in \mathscr{P} : d(P, P^*) > M\epsilon\}$ are complements of balls and hence typically not convex. A test function for non-convex alternatives can be constructed using a covering of $\mathscr{P}$ by convex sets. The entropy condition (3.7) controls the size of this cover and hence the rate of convergence in misspecified situations is determined by the covering numbers $N_t(\epsilon, \mathscr{P}, d; P_0, P^*)$. Because the validity of the theorem only relies on the existence of suitable tests, the entropy condition (3.7) could be replaced by a testing condition. To be precise: condition (3.7) can be replaced by the condition that the conclusion of theorem 3.11 is satisfied with $D(\epsilon) = e^{n\epsilon_n^2}$.

3.2.1 Distances and testing entropy

Because the entropies for testing are somewhat abstract, it is useful to relate them to ordinary entropy numbers. For our examples the bound given by the following lemma is useful. We assume that for some fixed constants $c, C > 0$ and for every $m \in \mathbb{N}$, $\lambda_1, \ldots, \lambda_m \geq 0$ with $\sum_i \lambda_i = 1$ and every $P, P_1, \ldots, P_m \in \mathscr{P}$ with $d(P, P_i) \leq c\, d(P, P^*)$,

$$\sum_i \lambda_i d^2(P_i, P^*) - C \sum_i \lambda_i d^2(P_i, P) \leq \sup_{0<\alpha<1} -\log P_0\Big(\frac{\sum_i \lambda_i p_i}{p^*}\Big)^\alpha. \tag{3.10}$$

Lemma 3.1. *If (3.10) holds, then there exists a constant $A > 0$ depending only on C such that for all $\epsilon > 0$, $N_t(\epsilon, \mathscr{P}, d; P_0) \leq N\big(A\epsilon, \{P \in \mathscr{P} : \epsilon < d(P, P^*) < 2\epsilon\}, d\big)$. (Any constant $A \leq (1/8) \wedge (1/4\sqrt{C})$ works.)*

Proof For a given constant $A > 0$ we can cover the set $\mathscr{P}_\epsilon := \{P \in \mathscr{P} : \epsilon < d(P, P^*) < 2\epsilon\}$ with $N = N\big(A\epsilon, \mathscr{P}_\epsilon, d\big)$ balls of radius $A\epsilon$. If the centers of these balls are not contained in $\mathscr{P}_\epsilon$, then we can replace these N balls by N balls of radius $2A\epsilon$ with centers in $\mathscr{P}_\epsilon$ whose union also covers the set $\mathscr{P}_\epsilon$. It suffices to show that (3.4) is valid for B_i equal to the convex hull of a typical ball B in this cover. Choose $2A < c$. If $P \in \mathscr{P}_\epsilon$ is the center of B and $P_i \in B$ for every i, then $d(P_i, P^*) \geq d(P, P^*) - 2A\epsilon$ by the triangle inequality and hence by assumption (3.10) the left side of (3.4) with $B_i = \mathrm{co}(B)$ is bounded below by $\sum_i \lambda_i\big((\epsilon - 2A\epsilon)^2 - C(2A\epsilon)^2\big)$. This is bounded below by $\epsilon^2/4$ for sufficiently small A. □

The logarithms $\log N\big(A\epsilon, \{P \in \mathscr{P} : \epsilon < d(P, P^*) < 2\epsilon\}, d\big)$ of the covering numbers in the preceding lemma are called 'local entropy numbers' and also the *Le Cam dimension* of the model $\mathscr{P}$ relative to the semi-metric d. They are bounded above by the simpler ordinary entropy numbers $\log N(A\epsilon, \mathscr{P}, d)$. The preceding lemma shows that the entropy condition (3.7) can be replaced by the ordinary entropy condition $\log N(\epsilon_n, \mathscr{P}, d) \leq n\epsilon_n^2$ whenever the semi-metric d satisfies (3.10).

If we evaluate (3.10) with $m = 1$ and $P_1 = P$, then we obtain, for every $P \in \mathscr{P}$,

$$d^2(P, P^*) \leq \sup_{0<\alpha<1} -\log P_0\Big(\frac{p}{p^*}\Big)^\alpha. \tag{3.11}$$

(Up to a factor 16 this inequality is also implied by finiteness of the covering numbers for testing.) This simpler condition gives an indication about the metrics d that may be used in combination with ordinary entropy. In lemma 3.2 we show that if d and the model $\mathscr{P}$ are convex, then the simpler condition (3.11) is equivalent to (3.10).

Because $-\log x \geq 1 - x$ for every $x > 0$, we can further simplify by bounding minus the logarithm in the right side by $1 - P_0(p/p^*)^\alpha$. This yields the bound

$$d^2(P, P^*) \leq \sup_{0<\alpha<1} \Big[1 - P_0\Big(\frac{p}{p^*}\Big)^\alpha\Big].$$

In the well-specified situation we have $P_0 = P^*$ and the right side for $\alpha = 1/2$ becomes $1 - \int \sqrt{p}\sqrt{p_0}\, d\mu$, which is $1/2$ times the Hellinger distance between P and P_0. In misspecified

situations this method of lower bounding can be useless, as $1 - P_0(p/p^*)^\alpha$ may be negative for $\alpha = 1/2$. On the other hand, a small value of α may be appropriate, as it can be shown that as $\alpha \downarrow 0$ the expression $1 - P_0(p/p^*)^\alpha$ is proportional to the difference of Kullback-Leibler divergences $P_0 \log(p^*/p)$, which is positive by the definition of P^*. If this approximation can be made uniform in p, then a semi-metric d which is bounded above by the Kullback-Leibler divergence can be used in the main theorem. We discuss this further in section 3.6 and use this in the examples of sections 3.4 and 3.5.

The case of convex models $\mathscr{P}$ is of interest, in particular for non- or semiparametric models and permits some simplification. For a convex model the point of minimal Kullback-Leibler divergence (if it exists) is automatically unique (up to redefinition on a null-set of P_0). Moreover, the expectations $P_0(p/p^*)$ are automatically finite, as required in theorem 3.1, and condition (3.10) is satisfied for a weighted Hellinger metric. We show this in lemma 3.3, after first showing that validity of the simpler lower bound (3.11) on the convex hull of $\mathscr{P}$ (if the semi-metric d is defined on this convex hull) implies the bound (3.10).

Lemma 3.2. *If d is defined on the convex hull of $\mathscr{P}$, the maps $P \mapsto d^2(P, P')$ are convex on $\mathrm{co}(\mathscr{P})$ for every $P' \in \mathscr{P}$ and the inequality (3.11) is valid for every P in the convex hull of $\mathscr{P}$, then (3.10) is satisfied for $\frac{1}{2}d$ instead of d.*

Lemma 3.3. *If $\mathscr{P}$ is convex and $P^* \in \mathscr{P}$ is a point at minimal Kullback-Leibler divergence with respect to P_0, then $P_0(p/p^*) \leq 1$ for every $P \in \mathscr{P}$ and (3.10) is satisfied with*

$$d^2(P_1, P_2) = \frac{1}{4} \int (\sqrt{p_1} - \sqrt{p_2})^2 \frac{p_0}{p^*}\, d\mu.$$

Proof For the proof of lemma 3.2 we first apply the triangle inequality repeatedly to find

$$\begin{aligned}
\sum_i \lambda_i d^2(P_i, P^*) &\leq 2 \sum_i \lambda_i d^2(P_i, P) + 2d^2(P, P^*) \\
&\leq 2 \sum_i \lambda_i d^2(P_i, P) + 4d^2\Big(P, \sum_i \lambda_i P_i\Big) + 4d^2\Big(\sum_i \lambda_i P_i, P^*\Big) \\
&\leq 6 \sum \lambda_i d^2(P_i, P) + 4d^2\Big(\sum_i \lambda_i P_i, P^*\Big),
\end{aligned}$$

by the convexity of d^2. It follows that

$$d^2(\sum_i \lambda_i P_i, P^*) \geq (1/4) \sum_i \lambda_i d^2(P_i, P^*) - 3/2 \sum_i \lambda_i d^2(P_i, P).$$

If (3.11) holds for $P = \sum_i \lambda_i P_i$, then we obtain (3.10) with d^2 replaced by $d^2/4$ and $C = 6$.

Next we prove lemma 3.3. For $P \in \mathscr{P}$ define a family of convex combinations $\{P_\lambda : \lambda \in [0,1]\} \subset \mathscr{P}$ by $P_\lambda = \lambda P + (1-\lambda)P^*$. For all values of $\lambda \in [0,1]$:

$$0 \leq f(\lambda) := -P_0 \log \frac{p_\lambda}{p^*} = -P_0 \log\Big(1 + \lambda\Big(\frac{p}{p^*} - 1\Big)\Big), \tag{3.12}$$

since $P^* \in \mathscr{P}$ is at minimal Kullback-Leibler divergence with respect to P_0 in $\mathscr{P}$ by assumption. For every fixed $y \geq 0$ the function $\lambda \mapsto \log(1 + \lambda y)/\lambda$ is non-negative and increases

monotonously to y as $\lambda \downarrow 0$. The function is bounded in absolute value by 2 for $y \in [-1, 0]$ and $\lambda \leq \frac{1}{2}$. Therefore, by the monotone and dominated convergence theorems applied to the positive and negative parts of the integrand in the right side of (3.12):

$$f'(0+) = 1 - P_0\Big(\frac{p}{p^*}\Big).$$

Combining the fact that $f(0) = 0$ with (3.12), we see that $f'(0+) \geq 0$ and hence we find $P_0(p/p^*) \leq 1$. The first assertion of lemma 3.3 now follows.

For the proof that (3.11) is satisfied, we first note that $-\log x \geq 1 - x$, so that it suffices to show that $1 - P_0(p/p^*)^{1/2} \geq d^2(P, P^*)$. Now

$$\int (\sqrt{p^*} - \sqrt{p})^2 \frac{p_0}{p^*}\, d\mu = 1 + P_0\frac{p}{p^*} - 2P_0\sqrt{\frac{p}{p^*}} \leq 2 - 2P_0\sqrt{\frac{p}{p^*}},$$

by the first part of the proof. □

3.2.2 Extensions

In this section we give some generalizations of theorem 3.1. Theorem 3.2 enables us to prove that optimal rates are achieved in parametric models. Theorem 3.3 extends theorem 3.1 to situations in which the model, the prior and the point P^* are dependent on n. Third, we consider the case in which the priors Π_n assign a mass slightly less than 1 to the models $\mathscr{P}_n$.

Theorem 3.1 does not give the optimal ($\sqrt{n}$-) rate of convergence for finite-dimensional models $\mathscr{P}$, both because the choice $\epsilon_n = 1/\sqrt{n}$ is excluded (by the condition $n\epsilon_n^2 \to \infty$) and because the prior mass condition is too restrictive. The following theorem remedies this, but is more complicated. The adapted prior mass condition takes the following form: for all natural numbers n and j,

$$\frac{\Pi\big(P \in \mathscr{P} : j\epsilon_n < d(P, P^*) < 2j\epsilon_n\big)}{\Pi\big(B(\epsilon_n, P^*; P_0)\big)} \leq e^{n\epsilon_n^2 j^2/8}. \tag{3.13}$$

Theorem 3.2. *For a given model $\mathscr{P}$ with prior Π and some $P^* \in \mathscr{P}$, assume that the Kullback-Leibler divergence $-P_0 \log(p^*/p_0)$ is finite and that $P_0(p/p^*) < \infty$ for all $P \in \mathscr{P}$. If ϵ_n are strictly positive numbers with $\epsilon_n \to 0$ and $\liminf n\epsilon_n^2 > 0$, such that (3.13) and (3.7) are satisfied, then, for every sequence $M_n \to \infty$, as $n \to \infty$,*

$$\Pi_n\big(\, P \in \mathscr{P} : d(P, P^*) \geq M_n\epsilon_n \,\big|\, X_1, \ldots, X_n \,\big) \to 0, \qquad \text{in } L_1(P_0). \tag{3.14}$$

There appears to be no compelling reason to choose the model $\mathscr{P}$ and the prior Π the same for every n. The validity of the preceding theorems does not depend on this. We formalize this fact in the following theorem. For each n, we let $\mathscr{P}_n$ be a set of probability measures on $(\mathscr{X}, \mathscr{A})$ given by densities p_n relative to a σ-finite measure μ_n on this space. Given a prior measure Π_n on an appropriate σ-field, we define the posterior by (3.1) with P^* and Π replaced by P_n^* and Π_n.

Theorem 3.3. *The preceding theorems remain valid if $\mathscr{P}$, Π, P^* and d depend on n, but satisfy the given conditions for each n (for a single constant L).*

As a final extension we note that the assertion

$$P_0^n \Pi_n(P \in \mathscr{P}_n : d_n(P, P_n^*) \geq M_n \epsilon_n \big| X_1, \ldots, X_n) \to 0$$

of the preceding theorems remains valid even if the priors Π_n do not put *all* their mass on the 'models' $\mathscr{P}_n$ (but the models $\mathscr{P}_n$ do satisfy the entropy condition). Of course, in such cases the posterior puts mass outside the model and it is desirable to complement the above assertion with the assertion that $\Pi_n(\mathscr{P}_n \big| X_1, \ldots, X_n) \to 1$ in $L_1(P_0)$. The latter is certainly true if the priors put only very small fractions of their mass outside the models $\mathscr{P}_n$. More precisely, the latter assertion is true if

$$\frac{1}{\Pi_n\big(B(\epsilon_n, P_n^*, P_0)\big)} \int_{\mathscr{P}_n^c} \Big(P \frac{p_0}{p_n^*}\Big)^n d\Pi_n(P) \leq o(e^{-2n\epsilon_n^2}). \tag{3.15}$$

This observation is not relevant for the examples in the present paper. However, it may prove relevant to alleviate the entropy conditions in the preceding theorems in certain situations. These conditions limit the complexity of the models and it seems reasonable to allow a trade-off between complexity and prior mass. Condition (3.15) allows a crude form of such a trade-off: a small part $\mathscr{P}_n^c$ of the model may be more complex, provided that it receives a negligible amount of prior mass.

3.2.3 Consistency

The preceding theorems yield a rate of convergence $\epsilon_n \to 0$, expressed as a function of prior mass and model entropy. In certain situations the prior mass and entropy may be hard to quantify. In contrast, for inferring consistency of the posterior such quantification is unnecessary. This could be proved directly, as Schwartz (1965) [82] achieved in the well-specified situation, but it can also be inferred from the preceding rate theorems. (A direct proof might actually give the same theorem with a slightly bigger set $B(\epsilon, P^*; P_0)$.) We consider this for the situation of theorem 3.1 only.

Corollary 3.1. *For a given model $\mathscr{P}$ with prior Π and some $P^* \in \mathscr{P}$, assume that the Kullback-Leibler divergence $-P_0 \log(p^*/p_0)$ is finite and that $P_0(p/p^*) < \infty$ for all $P \in \mathscr{P}$. Suppose that for every $\epsilon > 0$*

$$\begin{aligned} \Pi\big(B(\epsilon, P^*; P_0)\big) &> 0, \\ N_t\big(\epsilon, \mathscr{P}, d; P_0, P^*\big) &< \infty, \end{aligned} \tag{3.16}$$

Then for every $\epsilon > 0$, as $n \to \infty$,

$$\Pi_n\big(P \in \mathscr{P} : d(P, P^*) \geq \epsilon \;\big|\; X_1, \ldots, X_n \big) \to 0, \qquad \text{in } L_1(P_0^n). \tag{3.17}$$

Proof Define functions f and g as follows:

$$f(\epsilon) = \Pi\big(B(\epsilon, P^*; P_0)\big), \qquad g(\epsilon) = N_t\big(\epsilon, \mathscr{P}, d; P_0, P^*\big).$$

We shall show that there exists a sequence $\epsilon_n \to 0$ such that $f(\epsilon_n) \geq e^{-n\epsilon_n^2}$ and $g(\epsilon_n) \leq e^{n\epsilon_n^2}$ for all sufficiently large n. This implies that the conditions of theorem 3.1 are satisfied for this choice of ϵ_n and hence the posterior converges with rate at least ϵ_n.

To show the existence of ϵ_n define functions h_n by:

$$h_n(\epsilon) = e^{-n\epsilon^2}\Big(g(\epsilon) + \frac{1}{f(\epsilon)}\Big).$$

This is well defined and finite by the assumptions and $h_n(\epsilon) \to 0$ as $n \to \infty$, for every fixed $\epsilon > 0$. Therefore, there exists $\epsilon_n \downarrow 0$ with $h_n(\epsilon_n) \to 0$ (*e.g.* fix $n_1 < n_2 < \cdots$ such that $h_n(1/k) \leq 1/k$ for all $n \geq n_k$; next define $\epsilon_n = 1/k$ for $n_k \leq n < n_{k+1}$). In particular, there exists an N such that $h_n(\epsilon_n) \leq 1$ for $n \geq N$. This implies that $f(\epsilon_n) \geq e^{-n\epsilon_n^2}$ and $g(\epsilon_n) \leq e^{n\epsilon_n^2}$ for every $n \geq N$. □

3.2.4 Multiple points of minimal Kullback-Leibler divergence

In this section we extend theorem 3.1 to the situation that there exists a set $\mathscr{P}^*$ of minimal Kullback-Leibler points.

Multiple minimal points can arise in two very different ways. First consider the situation where the true distribution P_0 and the elements of the model $\mathscr{P}$ possess different supports. Because the observations are sampled from P_0, they fall with probability one in the set where $p_0 > 0$ and hence the exact nature of the elements p of the model $\mathscr{P}$ on the set $\{p_0 = 0\}$ is irrelevant. Clearly multiple minima arise if the model contains multiple extensions of P^* to the set on which $p_0 = 0$. In this case the observations do not provide the means to distinguish between these extensions and consequently no such distinction can be made by the posterior either. Theorems 3.1 and 3.2 may apply under this type of non-uniqueness, as long as we fix one of the minimal points, and the assertion of the theorem will be true for any of the minimal points as soon as it is true for one of the minimal points. This follows because under the conditions of the theorem, $d(P_1^*, P_2^*) = 0$ whenever P_1^* and P_2^* agree on the set $p_0 > 0$, in view of (3.11).

Genuine multiple points of minimal Kullback-Leibler divergence may occur as well. For instance, one might fit a model consisting of normal distributions with means in $(-\infty, -1] \cup [1, \infty)$ and variance one, in a situation where the true distribution is normal with mean 0. The normal distributions with means -1 and 1 both have the minimal Kullback-Leibler divergence. This situation is somewhat artificial and we are not aware of more interesting examples in the nonparametric or semiparametric case that interests us most in the present paper. Nevertheless, because it appears that the situation might arise, we give a brief discussion of an extension of theorem 3.1.

Rather than to a single measure $P^* \in \mathscr{P}$ the extension refers to a finite subset $\mathscr{P}^* \subset \mathscr{P}$ of points at minimal Kullback-Leibler divergence. We give conditions under which the posterior distribution concentrates asymptotically near this set of points. We redefine our 'covering numbers for testing under misspecification' $N_t(\epsilon, \mathscr{P}, d; P_0, \mathscr{P}^*)$ as the minimal number N of convex sets $B_1, \ldots, B_N$ of probability measures on $(\mathscr{X}, \mathscr{A})$ needed to cover the set $\{P \in \mathscr{P} : \epsilon < d(P, \mathscr{P}^*) < 2\epsilon\}$ such that

$$\sup_{P^* \in \mathscr{P}^*} \inf_{P \in B_i} \sup_{0<\alpha<1} -\log P_0\Big(\frac{p}{p^*}\Big)^\alpha \geq \frac{\epsilon^2}{4}. \tag{3.18}$$

This roughly says that for every $P \in \mathscr{P}$ there exists some minimal point to which we can apply arguments as before.

Theorem 3.4. *For a given model $\mathscr{P}$, prior Π on $\mathscr{P}$ and some subset $\mathscr{P}^* \subset \mathscr{P}$ assume that $-P_0 \log(p^*/p_0) < \infty$ and $P_0(p/p^*) < \infty$ for all $P \in \mathscr{P}$ and $P^* \in \mathscr{P}^*$. Suppose that there exists a sequence of strictly positive numbers ϵ_n with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ and a constant $L > 0$, such that for all n and all $\epsilon > \epsilon_n$:*

$$\inf_{P^* \in \mathscr{P}^*} \Pi\big(B(\epsilon_n, P^*; P_0)\big) \geq e^{-Ln\epsilon_n^2}, \tag{3.19}$$

$$N_t\big(\epsilon, \mathscr{P}, d; P_0, \mathscr{P}^*\big) \leq e^{n\epsilon_n^2}. \tag{3.20}$$

Then for every sufficiently large constant $M > 0$, as $n \to \infty$,

$$\Pi_n\big(\, P \in \mathscr{P} : d(P, \mathscr{P}^*) \geq M\epsilon_n \;\big|\; X_1, \ldots, X_n \,\big) \to 0, \qquad \text{in } L_1(P_0^n). \tag{3.21}$$

3.3 Mixtures

Let μ denote the Lebesgue measure on $\mathbb{R}$. For each $z \in \mathbb{R}$ let $x \mapsto p(x|z)$ be a fixed μ-probability density on a measurable space $(\mathscr{X}, \mathscr{A})$ that depends measurably on (x, z) and for a distribution F on $\mathbb{R}$ define the μ-density:

$$p_F(x) = \int p(x|z)\, dF(z).$$

Let P_F be the corresponding probability measure. In this section we consider mixture models $\mathscr{P} = \{P_F : F \in \mathscr{F}\}$, where $\mathscr{F}$ is the set of all probability distributions on a given compact interval $[-M, M]$. We consider consistency for general mixtures and derive a rate of convergence in the special case that the family $p(\cdot|z)$ is the normal location family, *i.e.* with ϕ the standard normal density:

$$p_F(x) = \int \phi(x - z)\, dF(z). \tag{3.22}$$

The observations are an *i.i.d.* sample $X_1, \ldots, X_n$ drawn from a distribution P_0 on $(\mathscr{X}, \mathscr{A})$ with μ-density p_0 which is not necessarily of the mixture form. As a prior for F we use a Dirichlet process distribution (see Ferguson (1973, 1974) [32, 33]) on $\mathscr{F}$.

3.3.1 General mixtures

We say that the model is P_0-identifiable if for all pairs $F_1, F_2 \in \mathscr{F}$:

$$F_1 \neq F_2 \quad \Rightarrow \quad P_0(\, p_{F_1} \neq p_{F_2}\,) > 0.$$

Imposing this condition on the model excludes the first way in which non-uniqueness of P^* may occur (as discussed in subsection 3.2.4).

Lemma 3.4. *Assume that $-P_0 \log(p_F/p_0) < \infty$ for some $F \in \mathscr{F}$. If the map $z \mapsto p(x|z)$ is continuous for all x, then there exists an $F^* \in \mathscr{F}$ that minimizes $F \mapsto -P_0 \log(p_F/p_0)$ over $\mathscr{F}$. If the model is P_0-identifiable, then this F^* is unique.*

Proof If F_n is a sequence in $\mathscr{F}$ with $F_n \to F$ for the weak topology on $\mathscr{F}$, then $p_{F_n}(x) \to p_F(x)$ for every x, since the kernel is continuous in z (and hence also bounded as a result of the compactness of $[-M, M]$) and the portmanteau lemma. Consequently, $p_{F_n} \to p_F$ in $L_1(\mu)$ by Scheffé's lemma. It follows that the map $F \mapsto p_F$ from $\mathscr{F}$ into $L_1(\mu)$ is continuous for the weak topology on $\mathscr{F}$. The set $\mathscr{F}$ is compact for this topology, by Prohorov's theorem. The Kullback-Leibler divergence $p \mapsto -P_0 \log(p/p_0)$ is lower semi-continuous as a map from $L_1(\mu)$ to $\mathbb{R}$. Therefore, the map $F \mapsto -P_0 \log(p_F/p_0)$ is lower semi-continuous on the compactum $\mathscr{F}$ and hence attains its infimum on $\mathscr{F}$.

The map $p \mapsto -P_0 \log(p/p_0)$ is also convex. By the strict convexity of the function $x \mapsto -\log x$ we have, for any $\lambda \in (0,1)$:

$$-P_0 \log\Big(\frac{\lambda p_1 + (1-\lambda)p_2}{p_0}\Big) < -\lambda P_0 \log \frac{p_1}{p_0} - (1-\lambda) P_0 \log \frac{p_2}{p_0},$$

unless $P_0(\, p_1 = p_2\,) = 1$. This shows that the point of minimum of $F \mapsto P_0 \log(p_F/p_0)$ is unique if $\mathscr{F}$ is P_0-identifiable. □

Thus a minimal Kullback-Leibler point P_{F^*} exists and is unique under mild conditions on the kernel p. Because the model is convex, lemma 3.3 next shows that (3.11) is satisfied for the weighted Hellinger distance, whose square is equal to

$$d^2(P_{F_1}, P_{F_2}) = \tfrac{1}{2} \int (\sqrt{p_{F_1}} - \sqrt{p_{F_2}})^2 \frac{p_0}{p_{F^*}}\, d\mu. \tag{3.23}$$

If $p_0/p_{F^*} \in L_\infty(\mu)$, then this distance is bounded by the squared Hellinger distance H between the measures P_{F_1} and P_{F_2}.

Because $\mathscr{F}$ is compact for the weak topology and the map $F \mapsto p_F$ from $\mathscr{F}$ to $L_1(\mu)$ is continuous, the model $\mathscr{P} = \{P_F : F \in \mathscr{F}\}$ is compact relative to the total variation distance. Because the Hellinger and total variation distances define the same uniform structure, the model is also compact relative to the Hellinger distance and hence it is totally bounded, *i.e.* the covering numbers $N(\epsilon, \mathscr{P}, H)$ are finite for all ϵ. Combined with the result of the preceding paragraph and lemma's 3.2 and 3.3 this yields that the entropy condition of Corollary 3.1 is satisfied for d as in (3.23) if $p_0/p_{F^*} \in L_\infty(\mu)$ and we obtain the following theorem.

Theorem 3.5. *If $p_0/p_{F^*} \in L_\infty(\mu)$ and $\Pi\big(B(\epsilon, P_{F^*}; P_0)\big) > 0$ for every $\epsilon > 0$, then $\Pi_n\big(F : d(P_F, P_{F^*}) \geq \epsilon \;\big|\; X_1, \ldots, X_n\big) \to 0$ in $L_1(P_0^n)$ for d given by (3.23).*

3.3.2 Gaussian mixtures

Next we specialize to the situation where $p(x|z) = \phi(x - z)$ is a Gaussian convolution kernel and derive the rate of convergence. The Gaussian convolution model is well known to be P_0-identifiable if P_0 is Lebesgue absolutely continuous (see *e.g.* Pfanzagl (1988) [73]). Let d be defined as in (3.23). We assume that P_0 is such that $-P_0 \log(p_F/p_0)$ is finite for some F, so that there exists a minimal Kullback-Leibler point F^*, by lemma 3.4.

Lemma 3.5. *If for some constant $C_1 > 0$, $d(p_{F_1}, p_{F_2}) \leq C_1 H(p_{F_1}, p_{F_2})$, then the entropy condition:*

$$\log N(\epsilon_n, \mathscr{P}, d) \leq n\epsilon_n^2,$$

is satisfied for ϵ_n a large enough multiple of $\log n/\sqrt{n}$.

Proof Because the square of the Hellinger distance is bounded above by the L_1-norm the assumption implies that $d^2(P_{F_1}, P_{F_2}) \leq C_1^2 \| P_{F_1} - P_{F_2} \|_1$. Hence, for all $\epsilon > 0$, we have $N(C_1\epsilon, \mathscr{P}, d) \leq N(\epsilon^2, \mathscr{P}, \| \cdot \|_1)$. As a result of lemma 3.3 in Ghosal and Van der Vaart (2001) [40], there exists a constant $C_2 > 0$ such that:

$$\| P_{F_1} - P_{F_2} \|_1 \leq C_2^2 \| P_{F_1} - P_{F_2} \|_\infty \max\Big\{1, M, \sqrt{\log_+ \frac{1}{\| P_{F_1} - P_{F_2} \|_\infty}}\Big\}, \tag{3.24}$$

from which it follows that $N(C_2^2\epsilon \log(1/\epsilon)^{1/2}, \mathscr{P}, \| \cdot \|_1) \leq N(\epsilon, \mathscr{P}, \| \cdot \|_\infty)$ for small enough ϵ. With the help of lemma 3.2 in Ghosal and Van der Vaart (2001) [40], we see that there exists a constant $C_3 > 0$ such that for all $0 < \epsilon < e^{-1}$:

$$\log N(\epsilon, \mathscr{P}, \| \cdot \|_\infty) \leq C_3 \Big(\log \frac{1}{\epsilon}\Big)^2.$$

Combining all of the above, we note that for small enough $\epsilon > 0$:

$$\log N\Big(C_1 C_2 \epsilon^{1/2} \big(\log \frac{1}{\epsilon}\big)^{1/4}, \mathscr{P}, d\Big) \leq \log N(\epsilon, \mathscr{P}, \| \cdot \|_\infty) \leq C_3 \Big(\log \frac{1}{\epsilon}\Big)^2.$$

So if we can find a sequence ϵ_n such that for all $n \geq 1$, there exists an $\epsilon > 0$ such that:

$$C_1 C_2 \epsilon^{1/2} \Big(\log \frac{1}{\epsilon}\Big)^{1/4} \leq \epsilon_n, \quad \text{and} \quad C_3 \Big(\log \frac{1}{\epsilon}\Big)^2 \leq n\epsilon_n^2,$$

then we have demonstrated that:

$$\log N(\epsilon_n, \mathscr{P}, d) \leq \log N\Big(C_1 C_2 \epsilon^{1/2} \big(\log \frac{1}{\epsilon}\big)^{1/4}, \mathscr{P}, d\Big) \leq C_3 \Big(\log \frac{1}{\epsilon}\Big)^2 \leq n\epsilon_n^2.$$

One easily shows that this is the case for $\epsilon_n = \max\{C_1 C_2, C_3\}(\log n/\sqrt{n})$ (in which case we choose, for fixed n, $\epsilon = 1/n$), if n is taken large enough. □

We are now in a position to apply theorem 3.1. We consider, for given $M > 0$, the location mixtures (3.22) with the standard normal density ϕ as the kernel. We choose the prior Π equal to a Dirichlet prior on $\mathscr{F}$ specified by a finite base measure α with compact support and positive, continuous Lebesgue-density on $[-M, M]$.

Theorem 3.6. *Let P_0 be a distribution on $\mathbb{R}$ dominated by Lebesgue measure μ. Assume that $p_0/p_{F^*} \in L_\infty(\mu)$. Then the posterior distribution concentrates its mass around P_{F^*} asymptotically, at the rate $\log n/\sqrt{n}$ relative to the distance d on $\mathscr{P}$ given by (3.23).*

Proof The set of mixture densities p_F with $F \in \mathscr{F}$ is bounded above and below by the upper and lower envelope functions

$$U(x) = \phi(x+M)1_{\{x<-M\}} + \phi(x-M)1_{\{x>M\}} + \phi(0)1_{\{-M\le x\le M\}},$$

$$L(x) = \phi(x-M)1_{\{x<0\}} + \phi(x+M)1_{\{x\ge 0\}},$$

So for any $F \in \mathscr{F}$,

$$P_0\Big(\frac{p_F}{p_{F^*}}\Big) \le P_0\frac{U}{L}$$
$$\le \frac{\phi(0)}{\phi(2M)}\,P_0[-M,M] + P_0\big(e^{-2MX}1_{\{X<-M\}} + e^{2MX}1_{\{X>M\}}\big) < \infty,$$

because p_0 is essentially bounded by a multiple of p_{F^*} and P_{F^*} has sub-Gaussian tials. In view of lemmas 3.2 and 3.3 the covering number for testing $N_t(\epsilon, \mathscr{P}, d; P_0)$ in (3.7) is bounded above by the ordinary metric covering number $N(A\epsilon, \mathscr{P}, d)$, for some constant A. Then lemma 3.5 demonstrates that the entropy condition (3.7) is satisfied for ϵ_n a large multiple of $\log n/\sqrt{n}$.

It suffices to verify the prior mass condition (3.6). Let ϵ be given such that $0 < \epsilon < e^{-1}$. By lemma 3.2 in Ghosal and Van der Vaart (2001) [40], there exists a discrete distribution function $F' \in D[-M,M]$ supported on at most $N \le C_2\log(1/\epsilon)$ points $\{z_1, z_2, \ldots, z_N\} \subset [-M,M]$ such that $\| p_{F^*} - p_{F'} \|_\infty \le C_1\epsilon$, where $C_1, C_2 > 0$ are constants that depend on M only. We write: $F' = \sum_{j=1}^N p_j\delta_{z_j}$. Without loss of generality, we may assume that the set $\{z_j : j = 1, \ldots, N\}$ is 2ϵ-separated. Namely, if this is not the case, we may choose a maximal 2ϵ-separated subset of $\{z_j : j = 1, \ldots, N\}$ and shift the weights p_j to the nearest point in the subset. A discrete F'' obtained in this fashion satisfies $\| p_{F'} - p_{F''} \|_\infty \le 2\epsilon \| \phi' \|_\infty$. So by virtue of the triangle inequality and the fact that the derivative of the standard normal kernel ϕ is bounded, a given F' may be replaced by a 2ϵ-separated F'' if the constant C_1 is changed accordingly.

By lemma 3.3 in Ghosal and Van der Vaart (2001) [40] there exists a constant D_1 such that the L_1-norm of the difference satisfies:

$$\| P_{F^*} - P_{F'} \|_1 \le D_1\,\epsilon\,\Big(\log\frac{1}{\epsilon}\Big)^{1/2},$$

for small enough ϵ. Using lemma 3.6 in Ghosal and Van der Vaart (2001) [40], we note moreover that there exists a constant D_2 such that, for any $F \in \mathscr{F}$:

$$\| P_F - P_{F'} \|_1 \le D_2\Big(\epsilon + \sum_{j=1}^N \big|F[z_j-\epsilon, z_j+\epsilon] - p_j\big|\Big).$$

So there exists a constant $D > 0$ such that if F satisfies $\sum_{j=1}^N |F[z_j-\epsilon, z_j+\epsilon] - p_j| \le \epsilon$, then

$$\| P_F - P_{F^*} \|_1 \le D\,\epsilon\Big(\log\frac{1}{\epsilon}\Big)^{1/2}.$$

Let $Q(P)$ be the measure defined by $dQ(P) = (p_0/p_{F^*})\,dP$. The assumption that p_0/p_{F^*} is essentially bounded implies that there exists a constant $K > 0$ such that:

$$\| Q(P_{F_1}) - Q(P_{F_2}) \|_1 \leq K \| P_{F_1} - P_{F_2} \|_1 ,$$

for all $F_1, F_2 \in \mathscr{F}$. Since $Q(P_{F^*}) = P_0$, it follows that there exists a constant $D' > 0$ such that for small enough $\epsilon > 0$:

$$\Big\{ F \in \mathscr{F} : \sum_{j=1}^{N} |F[z_j - \epsilon, z_j + \epsilon] - p_j| \leq \epsilon \Big\}$$
$$\subset \Big\{ F \in \mathscr{F} : \| Q(P_F) - P_0 \|_1 \leq (D')^2 \epsilon \Big(\log \frac{1}{\epsilon}\Big)^{1/2} \Big\}.$$

We have that $dQ(P_F)/dP_0 = p_F/p_{F^*}$ and $P_0(p_F/p_{F^*}) \leq P_0(U/L) < \infty$. The Hellinger distance is bounded by the square root of the L_1-distance. Therefore, applying lemma 3.18 with $\eta = \eta(\epsilon) = D'\epsilon^{1/2}(\log(1/\epsilon))^{1/4}$, we see that the set of measures P_F with F in the set on the right side of the last display is contained in the set

$$\Big\{ P_F : F \in \mathscr{F}, -P_0 \log \frac{p_F}{p_{F^*}} \leq \zeta^2(\epsilon), P_0 \Big(\log \frac{p_F}{p_{F^*}}\Big)^2 \leq \zeta^2(\epsilon) \Big\} \subset B(\zeta(\epsilon), P_{F^*}; P_0),$$

where $\zeta(\epsilon) = D''\eta(\epsilon)\big(\log(1/\eta(\epsilon))\big) \leq D''D'\epsilon^{1/2}\big(\log(1/\epsilon)\big)^{5/4}$, for an appropriate constant D'', and small enough ϵ. It follows that

$$\Pi\big(B(\zeta(\epsilon), P_{F^*}; P_0)\big) \geq \Pi\Big\{ F \in \mathscr{F} : \sum_{j=1}^{N} |F[z_j - \epsilon, z_j + \epsilon] - p_j| \leq \epsilon \Big\}.$$

Following Ghosal *et al.* (2000) [39] (lemma 6.1) or lemma A.2 in Ghosal and Van der Vaart (2001) [40], we see that the prior measure at the right hand side of the previous display is lower bounded by:

$$c_1 \exp\big(-c_2 N \log(1/\epsilon)\big) \geq \exp\big(-L(\log(1/\epsilon))^2\big) \geq \exp\big(-L'(\log(1/\zeta(\epsilon)))^2\big),$$

where $c_1 > 1$, $c_2 > 0$ are constants and $L = C_2 c_2 > 0$. So if we can find a sequence ϵ_n such that for each sufficiently large n, there exists an $\epsilon > 0$ such that:

$$\epsilon_n \geq \zeta(\epsilon), \qquad n\epsilon_n^2 \geq \Big(\log \frac{1}{\zeta(\epsilon)}\Big)^2,$$

then $\Pi\big(B(\epsilon_n, P_{F^*}; P_0)\big) \geq \Pi\big(B(\zeta(\epsilon), P_{F^*}; P_0)\big) \geq \exp(-L' n\epsilon_n^2)$ and hence (3.6) is satisfied. One easily shows that for $\epsilon_n = \log n/\sqrt{n}$ and $\zeta(\epsilon) = 1/\sqrt{n}$, the two requirements are fulfilled for sufficiently large n. □

3.4 Regression

Let P_0 be the distribution of a random vector (X, Y) satisfying $Y = f_0(X) + e_0$ for independent random variables X and e_0 taking values in a measurable space $(\mathscr{X}, \mathscr{A})$ and in $\mathbb{R}$,

respectively and a measurable function $f_0 : \mathscr{X} \to \mathbb{R}$. The variables X and e_0 have given marginal distributions, which may be unknown, but are fixed throughout the following. The purpose is to estimate the regression function f_0 based on a random sample of variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ with the same distribution as (X, Y).

A Bayesian approach to this problem might start from the specification of a prior distribution on a given class $\mathscr{F}$ of measurable functions $f : \mathscr{X} \to \mathbb{R}$. If the distributions of X and e_0 are known, this is sufficient to determine a posterior. If these distributions are not known, then one might proceed to introduce priors for these unknowns as well. The approach we take here is to fix the distribution of e_0 to a normal or Laplace distribution, while aware of the fact that its true distribution may be different. We investigate the consequences of the resulting model misspecification. We shall show that misspecification of the error distribution does not have serious consequences for estimation of the regression function. In this sense a nonparametric Bayesian approach possesses the same robustness to misspecification as minimum contrast estimation using least squares or minimum absolute deviation. We shall also see that the use of the Laplace distribution requires no conditions on the tail of the distribution of the errors, whereas the normal distribution appears to give good results only if these tails are not too big. Thus the tail robustness of minimum absolute deviation versus the nonrobustness of the method of least squares also extends to Bayesian regression.

We build the posterior based on a regression model $Y = f(X)+e$ for X and e independent, as is the assumption on the true distribution of (X, Y). If we assume that the distribution P_X of X has a known form, then this distribution cancels out of the expression for the posterior on f. If, instead, we put independent priors on f and P_X respectively, then the prior on P_X would disappear upon marginalization of the posterior of (f, P_X) relative to f. Thus for investigating the posterior for f we may assume without loss of generality that the marginal distribution of X is known. It can be absorbed in the dominating measure μ for the model.

For $f \in \mathscr{F}$, let P_f be the distribution of the random variable (X, Y) satisfying $Y = f(X)+e$ for X and e independent variables, X having the same distribution as before and e possessing a given density p, possibly different from the density of the true error e_0. We shall consider the cases that p is normal or Laplace. Given a prior Π on $\mathscr{F}$, the posterior distribution for f is given by

$$B \mapsto \frac{\int_B \prod_{i=1}^n p\big(Y_i - f(X_i)\big)\, d\Pi(f)}{\int \prod_{i=1}^n p\big(Y_i - f(X_i)\big)\, d\Pi(f)}.$$

We shall show that this distribution concentrates near $f_0 + \mathrm{E}e_0$ in the case that p is a normal density and near $f_0 + \mathrm{median}(e_0)$ if p is Laplace, if these translates of the true regression function f_0 are contained in the model $\mathscr{F}$. If the prior is misspecified also in the sense that $f_0 + \mu \notin \mathscr{F}$ (where μ is the expectation or median of e_0), then under some conditions this remains true with f_0 replaced by a 'projection' f^* of f_0 on $\mathscr{F}$. In agreement with the notation in the rest of the paper we shall denote the true distribution of an observation (X, Y) by P_0 (stressing that, in general, P_0 is different from P_f with $f = 0$). The model $\mathscr{P}$ as in the statement of the main results is the set of all distributions P_f on $\mathscr{X} \times \mathbb{R}$ with $f \in \mathscr{F}$.

3.4.1 Normal regression

Suppose that the density p is the standard normal density $p(z) = (2\pi)^{-1/2}\exp(-\frac{1}{2}z^2)$. Then, with $\mu = \mathrm{E}e_0$,

$$\begin{aligned} \log\frac{p_f}{p_{f_0}}(X,Y) &= -\tfrac{1}{2}(f-f_0)^2(X) + e_0(f-f_0)(X), \\ -P_0\log\frac{p_f}{p_{f_0}} &= \tfrac{1}{2}P_0(f-f_0-\mu)^2 - \tfrac{1}{2}\mu^2. \end{aligned} \tag{3.25}$$

It follows that the Kullback-Leibler divergence $f \mapsto -P_0\log(p_f/p_0)$ is minimized for $f = f^* \in \mathscr{F}$ minimizing the map $f \mapsto P_0(f-f_0-\mu)^2$.

In particular, if $f_0 + \mu \in \mathscr{F}$, then the minimizer is $f_0 + \mu$ and $P_{f_0+\mu}$ is the point in the model that is closest to P_0 in the Kullback-Leibler sense. If also $\mu = 0$, then, even though the posterior on P_f will concentrate asymptotically near P_{f_0}, which is typically not equal to P_0, the induced posterior on f will concentrate near the true regression function f_0. This favorable property of Bayesian estimation is analogous to that of least squares estimators, also for non-normal error distributions.

If $f_0 + \mu$ is not contained in the model, then the posterior for f will in general not be consistent. We assume that there exists a unique $f^* \in \mathscr{F}$ that minimizes $f \mapsto P_0(f-f_0-\mu)^2$, as is the case, for instance, if $\mathscr{F}$ is a closed, convex subset of $L_2(P_0)$. Under some conditions we shall show that the posterior concentrates asymptotically near f^*. If $\mu = 0$, then f^* is the projection of f_0 into $\mathscr{F}$ and the posterior still behaves in a desirable manner. For simplicity of notation we assume that $\mathrm{E}_0 e_0 = 0$.

The following lemma shows that (3.10) is satisfied for a multiple of the $L_2(P_0)$-distance on $\mathscr{F}$.

Lemma 3.6. *Let $\mathscr{F}$ be a class of uniformly bounded functions $f : \mathscr{X} \to \mathbb{R}$ such that either $f_0 \in \mathscr{F}$ or $\mathscr{F}$ is convex and closed in $L_2(P_0)$. Assume that f_0 is uniformly bounded, that $\mathrm{E}_0 e_0 = 0$ and that $\mathrm{E}_0 e^{M|e_0|} < \infty$ for every $M > 0$. Then there exist positive constants C_1, C_2, C_3 such that, for all $m \in \mathbb{N}$, $f, f_1, \ldots, f_m \in \mathscr{F}$ and $\lambda_1, \ldots, \lambda_m \geq 0$ with $\sum_i \lambda_i = 1$,*

$$\begin{aligned} P_0\log\frac{p_f}{p_{f^*}} &\leq -\tfrac{1}{2}P_0(f-f^*)^2, \\ P_0\Big(\log\frac{p_{f^*}}{p_f}\Big)^2 &\leq C_1 P_0(f-f^*)^2, \\ \sup_{0<\alpha<1} -\log P_0\Big(\frac{\sum_i \lambda_i p_{f_i}}{p_{f^*}}\Big)^\alpha &\geq C_2\sum_i \lambda_i\Big(P_0(f_i-f^*)^2 - C_3 P_0(f-f_i)^2\Big). \end{aligned} \tag{3.26}$$

Proof We have

$$\log\frac{p_f}{p_{f^*}}(X,Y) = -\tfrac{1}{2}\big[(f_0-f)^2 - (f_0-f^*)^2\big](X) - e_0(f^*-f)(X). \tag{3.27}$$

The second term on the right has mean zero by assumption. The first term on the right has expectation $-\frac{1}{2}P_0(f^*-f)^2$ if $f_0 = f^*$, as is the case if $f_0 \in \mathscr{F}$. Furthermore, if $\mathscr{F}$ is convex the minimizing property of f^* implies that $P_0(f_0-f^*)(f^*-f) \geq 0$ for every $f \in \mathscr{F}$ and then

the expectation of the first term on the right is bounded above by $-\frac{1}{2}P_0(f^* - f)^2$. Therefore, in both cases (3.26) holds.

From (3.27) we also have, with M a uniform upper bound on $\mathscr{F}$ and f_0,

$$P_0\big(\log \frac{p_f}{p_{f^*}}\big)^2 \le P_0\big[(f^* - f)^2(2f_0 - f - f^*)^2\big] + 2P_0 e_0^2 P_0(f^* - f)^2,$$
$$P_0\big(\log \frac{p_f}{p_{f^*}}\big)^2\big(\frac{p_f}{p_{f^*}}\big)^\alpha \le P_0\big[(f^* - f)^2(2f_0 - f - f^*)^2 + 2e_0^2(f^* - f)^2\big]e^{2\alpha(M^2 + M|e_0|)}.$$

Both right sides can be further bounded by a constant times $P_0(f - f^*)^2$, where the constant dependens on M and the distribution of e_0 only.

In view of lemma 3.8 (below) with $p = p_{f^*}$ and $q_i = p_{f_i}$, we see that there exists a constant $C > 0$ depending on M only such that for all $\lambda_i \ge 0$ with $\sum_i \lambda_i = 1$,

$$\Big| 1 - P_0\Big(\frac{\sum_i \lambda_i p_{f_i}}{p_{f^*}}\Big)^\alpha - \alpha P_0 \log \frac{p_{f^*}}{\sum_i \lambda_i p_{f_i}} \Big| \le 2\alpha^2 C \sum_i \lambda_i P_0(f_i - f^*)^2. \tag{3.28}$$

By lemma 3.8 with $\alpha = 1$ and $p = p_f$ and similar arguments we also have that, for any $f \in \mathscr{F}$,

$$\Big| 1 - P_0\Big(\frac{\sum_i \lambda_i p_{f_i}}{p_f}\Big) - P_0 \log \frac{p_f}{\sum_i \lambda_i p_{f_i}} \Big| \le 2C \sum_i \lambda_i P_0(f_i - f)^2.$$

For $\lambda_i = 1$ this becomes

$$\Big| 1 - P_0\Big(\frac{p_{f_i}}{p_f}\Big) - P_0 \log \frac{p_f}{p_{f_i}} \Big| \le 2C P_0(f_i - f)^2.$$

Taking differences we obtain that

$$\Big| P_0 \log \frac{p_f}{\sum_i \lambda_i p_{f_i}} - \sum_i \lambda_i P_0 \log \frac{p_f}{p_{f_i}} \Big| \le 4C \sum_i \lambda_i P_0(f_i - f)^2.$$

By the additivity of the logarithm this inequality remains true if f on the left is replaced by f^*. Combine the resulting inequality with (3.28) to find that

$$\begin{aligned} 1 - P_0\Big(\frac{\sum_i \lambda_i p_{f_i}}{p_{f^*}}\Big)^\alpha &\ge \alpha \sum_i \lambda_i P_0 \log \frac{p_{f^*}}{p_{f_i}} \\ &\qquad - 2\alpha^2 C \sum_i \lambda_i P_0(f^* - f_i)^2 - 4C \sum_i \lambda_i P_0(f_i - f)^2 \\ &\ge \Big(\frac{\alpha}{2} - 2\alpha^2 C\Big) \sum_i \lambda_i P_0(f^* - f_i)^2 - 4C \sum_i \lambda_i P_0(f_i - f)^2, \end{aligned}$$

where we have used (3.26). For sufficiently small $\alpha > 0$ and suitable constants C_2, C_3 the right side is bounded below by the right side of the lemma. Finally the left side of the lemma can be bounded by the supremum over $\alpha \in (0, 1)$ of the left side of the last display, since $-\log x \ge 1 - x$ for every $x > 0$. □

In view of the preceding lemma the estimation of the quantities involved in the main theorems can be based on the $L_2(P_0)$ distance.

The 'neighborhoods' $B(\epsilon, P_{f^*}; P_0)$ involved in the prior mass conditions (3.6) and (3.13) can be interpreted in the form

$$B(\epsilon, P_{f^*}; P_0) = \Big\{ f \in \mathscr{F} : P_0(f - f_0)^2 \leq P_0(f^* - f_0)^2 + \epsilon^2, P_0(f - f^*)^2 \leq \epsilon^2 \Big\}.$$

If $P_0(f - f^*)(f^* - f_0) = 0$ for every $f \in \mathscr{F}$ (as is the case if $f^* = f_0$ or if f^* lies in the interior of $\mathscr{F}$) then this reduces to an $L_2(P_0)$-ball around f^*, by Pythagoras' theorem.

In view of the preceding lemma and lemma 3.1 the entropy for testing in (3.7) can be replaced by the local entropy of $\mathscr{F}$ for the $L_2(P_0)$-metric. The rate of convergence of the posterior distribution guaranteed by theorem 3.1 is then also relative to the $L_2(P_0)$-distance. These observations yield the following theorem.

Theorem 3.7. *Assume the conditions of lemma 3.6 and in addition that $P_0(f-f^*)(f^*-f_0) = 0$ for every $f \in \mathscr{F}$. If ϵ_n is a sequence of strictly positive numbers with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ such that for a constant $L > 0$ and all n:*

$$\Pi\big(f \in \mathscr{F} : P_0(f - f^*)^2 \leq \epsilon_n^2\big) \geq e^{-Ln\epsilon_n^2}, \tag{3.29}$$

$$N\big(\epsilon_n, \mathscr{F}, \|\cdot\|_{P_0,2}\big) \leq e^{n\epsilon_n^2}, \tag{3.30}$$

then $\Pi_n\big(f \in \mathscr{F} : P_0(f - f^*)^2 \geq M\epsilon_n^2 \;\big|\; X_1, \ldots, X_n\big) \to 0$ in $L_1(P_0^n)$, for every sufficiently large constant M.

There are many special cases of interest of this theorem and the more general results that can be obtained from theorems 3.1 and 3.2 using the preceding reasoning. Some of these are considered in the context of the well-specified regression model (see Shen and Wasserman (2001) [83]). The necessary estimates on the prior mass and the entropy are not different for other problems than the regression model. Entropy estimates can also be found in work on rates of convergence of minimum contrast estimators. For these reasons we exclude a discussion of concrete examples.

The following pair of lemmas were used in the proof of the preceding results.

Lemma 3.7. *There exists a universal constant C such that for any probability measure P_0 and any finite measures P and Q and any $0 < \alpha \leq 1$,*

$$\Big|1 - P_0\Big(\frac{q}{p}\Big)^\alpha - \alpha P_0 \log \frac{p}{q}\Big| \leq \alpha^2 C P_0\Big[\Big(\log \frac{p}{q}\Big)^2 \Big(\Big(\frac{q}{p}\Big)^\alpha 1_{\{q>p\}} + 1_{\{q\leq p\}}\Big)\Big].$$

Proof The function R defined by $R(x) = (e^x - 1 - x)/(x^2 e^x)$ for $x \geq 0$ and $R(x) = (e^x - 1 - x)/x^2$ for $x \leq 0$ is uniformly bounded on $\mathbb{R}$ by a constant C. We can write

$$P_0\Big(\frac{q}{p}\Big)^\alpha = 1 + \alpha P_0 \log \frac{q}{p} + P_0 R\Big(\alpha \log \frac{q}{p}\Big)\Big(\alpha \log \frac{q}{p}\Big)^2 \Big[\Big(\frac{q}{p}\Big)^\alpha 1_{\{q>p\}} + 1_{\{q\leq p\}}\Big].$$

The lemma follows. □

Lemma 3.8. *There exists a universal constant C such that for any probability measure P_0 and all finite measures $P, Q_1, \ldots, Q_m$ and constants $0 < \alpha \leq 1$, $\lambda_i \geq 0$ with $\sum_i \lambda_i = 1$*

$$\Big|1 - P_0\Big(\frac{\sum_i \lambda_i q_i}{p}\Big)^\alpha - \alpha P_0 \log \frac{p}{\sum_i \lambda_i q_i}\Big| \leq 2\alpha^2 C \sum_i \lambda_i P_0\Big(\log \frac{q_i}{p}\Big)^2\Big[\Big(\frac{q_i}{p}\Big)^2 + 1\Big].$$

Proof In view of lemma 3.7 with $q = \sum_i \lambda_i q_i$, it suffices to bound

$$P_0\Big[\Big(\log \frac{\sum_i \lambda_i q_i}{p}\Big)^2\Big(\Big(\frac{\sum_i \lambda_i q_i}{p}\Big)^\alpha 1_{\sum_i \lambda_i q_i > p} + 1_{\sum_i \lambda_i q_i \leq p}\Big)\Big],$$

by the right side of the lemma. We can replace α in the display by 2 and make the expression larger. Next we bound the two terms corresponding to the decomposition by indicators separately.

By the convexity of the map $x \mapsto x \log x$

$$\Big(\log \frac{\sum_i \lambda_i q_i}{p}\Big)\Big(\frac{\sum_i \lambda_i q_i}{p}\Big) \leq \sum_i \lambda_i \Big(\log \frac{q_i}{p}\Big)\Big(\frac{q_i}{p}\Big).$$

If $\sum_i \lambda_i q_i > p$, then the left side is positive and the inequality is preserved when we square on both sides. Convexity of the map $x \mapsto x^2$ allows to bound the square of the right side as in the lemma.

By the concavity of the logarithm

$$-\log \frac{\sum_i \lambda_i q_i}{p} \leq -\sum_i \lambda_i \log \frac{q_i}{p}.$$

On the the set $\sum_i \lambda_i q_i < p$ the left side is positive and we can again take squares on both sides and preserve the inequality. □

3.4.2 Laplace regression

Suppose that the error-density p is equal to the Laplace density $p(x) = \frac{1}{2}\exp(-|x|)$. Then,

$$\log \frac{p_f}{p_{f_0}}(X, Y) = -\big(\big|e_0 + f_0(X) - f(X)\big| - |e_0|\big),$$
$$-P_0 \log \frac{p_f}{p_{f_0}} = P_0 \Phi(f - f_0),$$

for $\Phi(\nu) = \mathrm{E}_0\big(|e_0 - \nu| - |e_0|\big)$. The function Φ is minimized over $\nu \in \mathbb{R}$ at the median of e_0. It follows that if $f_0 + m$, for m the median of e_0, is contained in $\mathscr{F}$, then the Kullback-Leibler divergence $-P_0 \log(p_f/p_0)$ is minimized over $f \in \mathscr{F}$ at $f = f_0 + m$. If $\mathscr{F}$ is a compact, convex subset of $L_1(P_0)$, then in any case there exists $f^* \in \mathscr{F}$ that minimizes the Kullback-Leibler divergence, but it appears difficult to determine this concretely in general. For simplicity of notation we shall assume that $m = 0$.

If the distribution of e_0 is smooth, then the function Φ will be smooth too. Because it is minimal at $\nu = m = 0$ it is reasonable to expect that, for ν in a neighborhood of $m = 0$ and some positive constant C_0

$$\Phi(\nu) = \mathrm{E}_0\big(|e_0 - \nu| - |e_0|\big) \geq C_0|\nu|^2. \tag{3.31}$$

Because Φ is convex, it is also reasonable to expect that its second derivative, if it exists, is strictly positive.

Lemma 3.9. *Let $\mathscr{F}$ be a class of uniformly bounded functions $f : \mathscr{X} \to \mathbb{R}$ and let f_0 be uniformly bounded. Assume that either $f_0 \in \mathscr{F}$ and (3.31) holds, or that $\mathscr{F}$ is convex and compact in $L_1(P_0)$ and that Φ is twice continuously differentiable with strictly positive second derivative. Then there exist positive constants C_0, C_1, C_2, C_3 such that, for all $m \in \mathbb{N}$, $f, f_1, \ldots, f_m \in \mathscr{F}$ and $\lambda_1, \ldots, \lambda_m \geq 0$ with $\sum_i \lambda_i = 1$,*

$$P_0 \log \frac{p_f}{p_{f^*}} \leq -C_0 P_0 (f - f^*)^2, \tag{3.32}$$

$$P_0 \Big(\log \frac{p_{f^*}}{p_f} \Big)^2 \leq C_1 P_0 (f - f^*)^2,$$

$$\sup_{0<\alpha<1} - \log P_0 \Big(\frac{\sum_i \lambda_i p_{f_i}}{p_{f^*}} \Big)^\alpha \geq C_2 \sum_i \lambda_i \Big(P_0 (f_i - f^*)^2 - C_3 P_0 (f - f_i)^2 \Big).$$

Proof Suppose first that $f_0 \in \mathscr{F}$, so that $f^* = f_0$. As Φ is monotone on $(0, \infty)$ and $(-\infty, 0)$, inequality (3.31) is automatically also satisfied for ν in a given compactum (with C_0 depending on the compactum). Choosing the compactum large enough such that $(f - f^*)(X)$ is contained in it with probability one, we conclude that (3.32) holds (with $f_0 = f^*$).

If f^* is not contained in $\mathscr{F}$ but $\mathscr{F}$ is convex, we obtain a similar inequality with f^* replacing f_0, as follows. Because f^* minimizes $f \mapsto P_0 \Phi(f - f_0)$ over $\mathscr{F}$ and $f_t = (1-t)f^* + tf \in \mathscr{F}$ for $t \in [0, 1]$, the right derivative of the map $t \mapsto P_0 \Phi(f_t - f_0)$ is nonnegative at $t = 0$. This yields $P_0 \Phi'(f^* - f_0)(f - f^*) \geq 0$. By a Taylor expansion

$$\begin{aligned} P_0 \log \frac{p_{f^*}}{p_f} &= P_0 \big(\Phi(f - f_0) - \Phi(f^* - f_0) \big) \\ &= P_0 \Phi'(f^* - f_0)(f - f^*) + \tfrac{1}{2} P_0 \Phi''(\tilde{f} - f_0)(f - f^*)^2, \end{aligned}$$

for some $\tilde{f}$ between f and f^*. The first term on the right is nonnegative and the function Φ'' is bounded away from zero on compacta by assumption. Thus the right side is bounded below by a constant times $P_0(f - f^*)^2$ and again (3.32) follows.

Because $\log(p_f / p_{f^*})$ is bounded in absolute value by $|f - f^*|$, we also have, with M a uniform upper bound on $\mathscr{F}$ and f_0,

$$P_0 \Big(\log \frac{p_f}{p_{f^*}} \Big)^2 \leq P_0 (f^* - f)^2,$$

$$P_0 \Big(\log \frac{p_f}{p_{f^*}} \Big)^2 \Big(\frac{p_f}{p_{f^*}} \Big)^\alpha \leq P_0 (f^* - f)^2 e^{2\alpha M}.$$

As in the proof of lemma 3.6 we can combine these inequalities, (3.32) and lemma 3.8 to obtain the result. □

As for regression using the Laplace density for the error-distribution, the preceding lemma reduces the entropy calculations for the application of theorem 3.1 to estimates of the $L_2(P_0)$-entropy of the class of regression functions $\mathscr{F}$. The resulting rate of convergence is the same

as in the case where a normal distribution is used for the error. A difference with the normal case is that presently no tail conditions of the type $\mathrm{E}_0 e^{\epsilon|e_0|} < \infty$ are necessary. Instead the lemma assumes a certain smoothness of the true distribution of the error e_0.

3.5 Parametric models

The behavior of posterior distributions for finite-dimensional, misspecified models was considered in Berk (1966) [8] and more recently by Bunke and Milhaud (1998) [20] (see also the references in the latter). In this section we show that the basic result that the posterior concentrates near a minimal Kullback-Leibler point at the rate $\sqrt{n}$ follows from our general theorems under some natural conditions. We first consider models indexed by a parameter in a general metric space and relate the rate of convergence to the metric entropy of the parameter set. Next we specialize to Euclidean parameter sets.

Let $\{p_\theta : \theta \in \Theta\}$ be a collection of probability densities indexed by a parameter θ in a metric space (Θ, d). Let P_0 be the true distribution of the data and assume that there exists a $\theta^* \in \Theta$, such that for all $\theta, \theta_1, \theta_2 \in \Theta$ and some constant $C > 0$,

$$P_0 \log \frac{p_\theta}{p_{\theta^*}} \leq -C\, d^2(\theta, \theta^*), \tag{3.33}$$

$$P_0 \Big(\sqrt{\frac{p_{\theta_1}}{p_{\theta_2}}} - 1 \Big)^2 \leq d^2(\theta_1, \theta_2), \tag{3.34}$$

$$P_0 \Big(\log \frac{p_{\theta_1}}{p_{\theta_2}} \Big)^2 \leq d^2(\theta_1, \theta_2). \tag{3.35}$$

The first inequality implies that θ^* is a point of minimal Kullback-Leibler divergence $\theta \mapsto -P_0 \log(p_\theta / p_0)$ between P_0 and the model. The second and third conditions are (integrated) Lipschitz conditions on the dependence of p_θ on θ. The following lemma shows that in the application of theorems 3.1 and 3.2 these conditions allow one to replace the entropy for testing by the local entropy of Θ relative to (a multiple of) the natural metric d .

In examples it may be worthwhile to relax the conditions somewhat. In particular, the conditions (3.34)-(3.35) can be 'localized'. Rather than assuming that they are valid for every $\theta_1, \theta_2 \in \Theta$ the same results can be obtained if they are valid for every pair (θ_1, θ_2) with $d(\theta_1, \theta_2)$ sufficiently small and every pair (θ_1, θ_2) with arbitrary θ_1 and $\theta_2 = \theta^*$. For $\theta_2 = \theta^*$ and $P_0 = P_{\theta^*}$ (*i.e.* the well-specified situation), condition (3.34) is a bound on the Hellinger distance between P_{θ^*} and P_{θ_1}.

Lemma 3.10. *Under the preceding conditions there exist positive constants C_1, C_2 such that, for all $m \in \mathbb{N}, \theta, \theta_1, \ldots, \theta_m \in \Theta$ and $\lambda_1, \ldots, \lambda_m \geq 0$ with $\sum_i \lambda_i = 1$,*

$$\sum_i \lambda_i d^2(\theta_i, \theta^*) - C_1 \sum_i \lambda_i d^2(\theta, \theta_i) \leq C_2 \sup_{0<\alpha<1} -\log P_0 \Big(\frac{\sum_i \lambda_i p_{\theta_i}}{p_{\theta^*}} \Big)^\alpha.$$

Proof In view of lemma 3.12 (below) with $p = p_{\theta^*}$, (3.34) and (3.35), there exists a constant

C such that

$$\Big| 1 - P_0\Big(\frac{\sum_i \lambda_i p_{\theta_i}}{p_{\theta^*}}\Big)^\alpha - \alpha P_0\Big(\log \frac{p_{\theta^*}}{\sum_i \lambda_i p_{\theta_i}}\Big) \Big| \le 2\alpha^2 C \sum_i \lambda_i d^2(\theta_i, \theta^*). \tag{3.36}$$

By lemma 3.12 with $\alpha = 1$, $p = p_\theta$, (3.34) and (3.35),

$$\Big| 1 - P_0\Big(\frac{\sum_i \lambda_i p_{\theta_i}}{p_\theta}\Big) - P_0\Big(\log \frac{p_\theta}{\sum_i \lambda_i p_{\theta_i}}\Big) \Big| \le 2C \sum_i \lambda_i d^2(\theta_i, \theta).$$

We can evaluate this with $\lambda_i = 1$ (for each i in turn) and next subtract the convex combination of the resulting inequalities from the preceding display to obtain

$$\Big| P_0\Big(\log \frac{p_\theta}{\sum_i \lambda_i p_{\theta_i}}\Big) - \sum_i \lambda_i P_0\Big(\log \frac{p_\theta}{p_{\theta_i}}\Big) \Big| \le 4C \sum_i \lambda_i d^2(\theta_i, \theta).$$

By the additivity of the logarithm this remains valid if we replace θ in the left side by θ^*. Combining the resulting inequality with (3.33) and (3.36) we obtain

$$1 - P_0\Big(\frac{\sum_i \lambda_i p_{\theta_i}}{p_{\theta^*}}\Big)^\alpha \ge \alpha \sum_i \lambda_i d^2(\theta_i, \theta^*)(C - 2\alpha) - 4C \sum_i \lambda_i d^2(\theta_i, \theta).$$

The lemma follows upon choosing $\alpha > 0$ sufficiently small and using $-\log x \ge 1 - x$. □

If the prior on the model $\{p_\theta : \theta \in \Theta\}$ is induced by a prior on the parameter set Θ, then the prior mass condition (3.13) translates into a lower bound for the prior mass of the set

$$B(\epsilon, \theta^*; P_0) = \Big\{\theta \in \Theta : -P_0 \log \frac{p_\theta}{p_{\theta^*}} \le \epsilon^2, P_0\Big(\log \frac{p_\theta}{p_{\theta^*}}\Big)^2 \le \epsilon^2\Big\}.$$

In addition to (3.33), it is reasonable to assume a lower bound of the form

$$P_0 \log \frac{p_\theta}{p_{\theta^*}} \ge -\underline{C}\, d^2(\theta, \theta^*), \tag{3.37}$$

at least for small values of $d(\theta, \theta^*)$. This together with (3.35) implies that $B(\epsilon, \theta^*; P_0)$ contains a ball of the form $\{\theta : d(\theta, \theta^*) \le C_1\epsilon\}$ for small enough ϵ. Thus in the verification of (3.6) or (3.13) we may replace $B(\epsilon, P^*; P_0)$ by a ball of radius ϵ around θ^*. These observations lead to the following theorem.

Theorem 3.8. *Let (3.33)–(3.37) hold. If for sufficiently small A and C,*

$$\sup_{\epsilon > \epsilon_n} \log N(A\epsilon, \{\theta \in \Theta : \epsilon < d(\theta, \theta^*) < 2\epsilon\}, d) \le n\epsilon_n^2,$$
$$\frac{\Pi(\theta : j\epsilon_n < d(\theta, \theta^*) < 2j\epsilon_n)}{\Pi(\theta : d(\theta, \theta^*) \le C\epsilon_n)} \le e^{n\epsilon_n^2 j^2/8},$$

then $\Pi\big(\theta : d(\theta, \theta^*) \ge M_n\epsilon_n \,|\, X_1, \ldots, X_n\big) \to 0$ in $L_1(P_0^n)$ for any $M_n \to \infty$.

3.5.1 Finite-dimensional models

Let Θ be an open subset of m-dimensional Euclidean space equipped with the Euclidean distance d and let $\{p_\theta : \theta \in \Theta\}$ be a model satisfying (3.33)–(3.37).

Then the local covering numbers as in the preceding theorem satisfy, for some constant B,

$$N\big(A\epsilon, \{\theta \in \Theta : \epsilon < d(\theta, \theta^*) < 2\epsilon\}, d\big) \leq \Big(\frac{B}{A}\Big)^m,$$

(see *e.g.* Ghosal *et al.* (2000) [39], section 5). In view of lemma 3.2, condition (3.7) is satisfied for ϵ_n a large multiple of $1/\sqrt{n}$. If the prior Π on Θ possesses a density that is bounded away from zero and infinity, then

$$\frac{\Pi\big(\theta : d(\theta, \theta^*) \leq j\epsilon\big)}{\Pi\big(B(\epsilon, \theta^*; P_0)\big)} \leq C_2 j^m,$$

for some constant C_2. It follows that (3.13) is satisfied for the same ϵ_n. Hence the posterior concentrates at rate $1/\sqrt{n}$ near the point θ^*.

The preceding situation arises if the minimal point θ^* is interior to the parameter set Θ. An example is fitting an exponential family, such as the Gaussian model, to observations that are not sampled from an element of the family. If the minimal point θ^* is not interior to Θ, then we cannot expect (3.33) to hold for the natural distance and different rates of convergence may arise. We include a simple example of the latter type, which is somewhat surprising.

Example 3.1. Suppose that P_0 is the standard normal distribution and the model consists of all $N(\theta, 1)$-distributions with $\theta \geq 1$. The minimal Kullback-Leibler point is $\theta^* = 1$. If the prior possesses a density on $[1, \infty)$ that is bounded away from 0 and infinity near 1, then the posterior concentrates near θ^* at the rate $1/n$.

One easily shows that:

$$\begin{aligned} -P_0 \log \frac{p_\theta}{p_{\theta^*}} &= \tfrac{1}{2}(\theta - \theta^*)(\theta + \theta^*), \\ -\log P_0 (p_\theta / p_{\theta^*})^\alpha &= \tfrac{1}{2}\alpha(\theta - \theta^*)\big(\theta + \theta^* - \alpha(\theta - \theta^*)\big). \end{aligned} \tag{3.38}$$

This shows that (3.11) is satisfied for a multiple of the metric $d(p_{\theta_1}, p_{\theta_2}) = \sqrt{|\theta_1 - \theta_2|}$ on $\Theta = [1, \infty)$. Its strengthening (3.10) can be verified by the same methods as before, or alternatively the existence of suitable tests can be established directly based on the special nature of the normal location family. (A suitable test for an interval (θ_1, θ_2) can be obtained from a suitable test for its left end point.) The entropy and prior mass can be estimated as in regular parametric models and conditions (3.7)-(3.13) can be shown to be satisfied for ϵ_n a large multiple of $1/\sqrt{n}$. This yields the rate $1/\sqrt{n}$ relative to the metric $\sqrt{|\theta_1 - \theta_2|}$ and hence the rate $1/n$ in the natural metric.

Theorem 3.2 only gives an upper bound on the rate of convergence. In the present situation this appears to be sharp. For instance, for a uniform prior on $[1, 2]$ the posterior mass of the

interval $[c, 2]$ can be seen to be, with $Z_n = \sqrt{n}\bar{X}_n$,

$$\frac{\Phi(2\sqrt{n} - Z_n) - \Phi(c\sqrt{n} - Z_n)}{\Phi(2\sqrt{n} - Z_n) - \Phi(\sqrt{n} - Z_n)} \approx \frac{\sqrt{n} - Z_n}{c\sqrt{n} - Z_n} e^{-\frac{1}{2}(c^2-1)n + Z_n(c-1)\sqrt{n}},$$

where we use Mills' ratio to see that $\Phi(y_n) - \Phi(x_n) \approx (1/x_n)\phi(x_n)$ if $x_n, y_n \to c \in (0, 1)$ such that $x_n/y_n \to 0$. This is bounded away from zero for $c = c_n = 1 + C/n$ and fixed C.

Lemma 3.11. *There exists a universal constant C such for any probability measure P_0 and any finite measures P and Q and any $0 < \alpha \leq 1$,*

$$\Big| 1 - P_0\Big(\frac{q}{p}\Big)^{\alpha} - \alpha P_0 \log \frac{p}{q} \Big| \leq \alpha^2 C P_0\Big[\Big(\sqrt{\frac{q}{p}} - 1\Big)^2 1_{\{q>p\}} + \Big(\log \frac{p}{q}\Big)^2 1_{\{q \leq p\}}\Big].$$

Lemma 3.12. *There exists a universal constant C such that, for any probability measure P_0 and any finite measures P, $Q_1, \ldots, Q_m$ and any $\lambda_1, \ldots, \lambda_m \geq 0$ with $\sum_i \lambda_i = 1$ and $0 < \alpha \leq 1$, the following inequality holds:*

$$\Big|1 - P_0\Big(\frac{\sum_i \lambda_i q_i}{p}\Big)^{\alpha} - \alpha P_0 \log \frac{p}{\sum_i \lambda_i q_i}\Big| \leq 2\alpha^2 C \sum_i \lambda_i P_0\Big[\Big(\sqrt{\frac{q_i}{p}} - 1\Big)^2 + \Big(\log \frac{q_i}{p}\Big)^2\Big].$$

Proofs The function R defined by $R(x) = (e^x - 1 - x)/\alpha^2(e^{x/2\alpha} - 1)^2$ for $x \geq 0$ and $R(x) = (e^x - 1 - x)/x^2$ for $x \leq 0$ is uniformly bounded on $\mathbb{R}$ by a constant C, independent of $\alpha \in (0, 1]$. (This may be proved by noting that the functions $(e^x - 1)/\alpha(e^{\alpha x} - 1)$ and $(e^x - 1 - x)/(e^{x/2} - 1)^2$ are bounded, where this follows for the first by developing the exponentials in their power series.) For the proof of the first lemma, we can proceed as in the proof of lemma 3.7. For the proof of the second lemma we proceed as in the proof of lemma 3.8, this time also making use of the convexity of the map $x \mapsto |\sqrt{x} - 1|^2$ on $[0, \infty)$. □

3.6 Existence of tests

The proofs of theorems 3.1 and 3.2 rely on tests of P_0 versus the positive, finite measures $Q(P)$ obtained from points P that are at positive distance from the set of points with minimal Kullback-Leibler divergence. Because we need to test P_0 against finite measures (*i.e.* not necessarily probability measures), known results on tests using the Hellinger distance, such as in Le Cam (1986) [67] or Ghosal *et al.* (2000) [39] do not apply. It turns out that in this situation the Hellinger distance may not be appropriate and instead we use the full Hellinger transform. The aim of this section is to prove the existence of suitable tests and give upper bounds on their power. We first formulate the results in a general notation and then specialize to the application in misspecified models.

3.6.1 General setup

Let P be a probability measure on a measurable space $(\mathscr{X}, \mathscr{A})$ (playing the role of P_0) and let $\mathscr{Q}$ be a class of finite measures on $(\mathscr{X}, \mathscr{A})$ (playing the role of the measures Q with

$dQ = (p_0/p^*)\,dP$). We wish to bound the minimax risk for testing P versus $\mathscr{Q}$, defined by

$$\pi(P, \mathscr{Q}) = \inf_{\phi} \sup_{Q \in \mathscr{Q}} \big(P\phi + Q(1-\phi) \big),$$

where the infimum is taken over all measurable functions $\phi : \mathscr{X} \to [0,1]$. Let $\mathrm{co}(\mathscr{Q})$ denote the convex hull of the set $\mathscr{Q}$.

Lemma 3.13. *If there exists a σ-finite measure that dominates all $Q \in \mathscr{Q}$, then:*

$$\pi(P, \mathscr{Q}) = \sup_{Q \in \mathrm{co}(\mathscr{Q})} \big(P(p < q) + Q(p \geq q) \big).$$

Moreover, there exists a test ϕ that attains the infimum in the definition of $\pi(P, \mathscr{Q})$.

Proof If μ' is a measure dominating $\mathscr{Q}$, then a σ-finite measure μ exists that dominates both $\mathscr{Q}$ and P (for instance, $\mu = \mu' + P$). Let p and q be μ-densities of P and Q, for every $Q \in \mathscr{Q}$. The set of test-functions ϕ can be identified with the positive unit ball Φ of $L_\infty(\mathscr{X}, \mathscr{A}, \mu)$, which is dual to $L_1(\mathscr{X}, \mathscr{A}, \mu)$, since μ is σ-finite. If equipped with the weak-$*$ topology, the positive unit ball Φ is Hausdorff and compact by the Banach-Alaoglu theorem (see *e.g.* Megginson (1998) [70], theorem 2.6.18). The convex hull $\mathrm{co}(\mathscr{Q})$ (or rather the corresponding set of μ-densities) is a convex subset of $L_1(\mathscr{X}, \mathscr{A}, \mu)$. The map:

$$\begin{aligned} L_\infty(\mathscr{X}, \mathscr{A}, \mu) \times L_1(\mathscr{X}, \mathscr{A}, \mu) &\to \mathbb{R}, \\ (\phi, Q) &\mapsto \phi P + (1-\phi)Q, \end{aligned}$$

is concave in Q and convex in ϕ. (Note that in the current context we write ϕP instead of $P\phi$, in accordance with the fact that we consider ϕ as a bounded linear functional on $L_1(\mathscr{X}, \mathscr{A}, \mu)$.) Moreover, the map is weak-$*$-continuous in ϕ for every fixed Q (note that every weak-$*$-converging net $\phi_\alpha \xrightarrow{\text{w}-*} \phi$ by definition satisfies $\phi_\alpha Q \to \phi Q$ for all $Q \in L_1(\mathscr{X}, \mathscr{A}, \mu)$). The conditions for application of the minimax theorem (see, *e.g.*, Strasser (1985) [85], p. 239) are satisfied and we conclude:

$$\inf_{\phi \in \Phi} \sup_{Q \in \mathrm{co}(\mathscr{Q})} \big(\phi P + (1-\phi)Q \big) = \sup_{Q \in \mathrm{co}(\mathscr{Q})} \inf_{\phi \in \Phi} \big(\phi P + (1-\phi)Q \big).$$

The expression on the left side is the minimax testing risk $\pi(P, \mathscr{Q})$. The infimum in the right side is attained at the point $\phi = 1\{p < q\}$, which leads to the first assertion of the lemma upon substitution.

The second assertion of the lemma follows because the function $\phi \mapsto \sup\{\phi P + (1-\phi)Q : Q \in \mathrm{co}(\mathscr{Q})\}$ is a supremum of weak-$*$-continuous functions and hence attains its min+imum on the compactum Φ. □

It is possible to express the right side of the preceding lemma in the L_1-distance between P and Q, but this is not useful for the following. Instead, we use a bound in terms of the *Hellinger transform* $\rho_\alpha(P, Q)$, defined by, for $0 < \alpha < 1$,

$$\rho_\alpha(P, Q) = \int p^\alpha q^{1-\alpha}\, d\mu.$$

By Hölder's inequality this quantity is finite for all finite measures P and Q. The definition is independent of the choice of dominating measure μ.

For any pair (P, Q) and every $\alpha \in (0, 1)$, we can bound

$$\begin{aligned} P(p < q) + Q(p \geq q) &= \int_{p<q} p \, d\mu + \int_{p \geq q} q \, d\mu \\ &\leq \int_{p<q} p^{\alpha} q^{1-\alpha} \, d\mu + \int_{p \geq q} p^{\alpha} q^{1-\alpha} \, d\mu = \rho_{\alpha}(P, Q). \end{aligned} \tag{3.39}$$

Hence the right side of the preceding lemma is bounded by $\sup_Q \rho_\alpha(P, Q)$, for all $\alpha \in (0, 1)$. The advantage of this bound is the fact that it factorizes if P and Q are product measures. For ease of notation define

$$\rho_{\alpha}(\mathscr{P}, \mathscr{Q}) = \sup\bigl\{\rho_{\alpha}(P, Q) : P \in \mathrm{co}(\mathscr{P}), Q \in \mathrm{co}(\mathscr{Q})\bigr\}.$$

Lemma 3.14. *For any $0 < \alpha < 1$ and classes $\mathscr{P}_1$, $\mathscr{P}_2$, $\mathscr{Q}_1$, $\mathscr{Q}_2$ of finite measures:*

$$\rho_{\alpha}(\mathscr{P}_1 \times \mathscr{P}_2, \mathscr{Q}_1 \times \mathscr{Q}_2) \leq \rho_{\alpha}(\mathscr{P}_1, \mathscr{Q}_1) \, \rho_{\alpha}(\mathscr{P}_2, \mathscr{Q}_2),$$

where $\mathscr{P}_1 \times \mathscr{P}_2$ denotes the class of product measures $\{P_1 \times P_2 : P_1 \in \mathscr{P}_1, P_2 \in \mathscr{P}_2\}$.

Proof Let $P \in \mathrm{co}(\mathscr{P}_1 \times \mathscr{P}_2)$ and $Q \in \mathrm{co}(\mathscr{Q}_1 \times \mathscr{Q}_2)$ be given. Since both are (finite) convex combinations, σ-finite measures μ_1 and μ_2 can always be found such that both P and Q have $\mu_1 \times \mu_2$ densities which both can be written in the form of a finite convex combination as follows:

$$\begin{aligned} p(x, y) &= \sum_i \lambda_i p_{1i}(x) p_{2i}(y), \quad \lambda_i \geq 0, \quad \sum_i \lambda_i = 1, \\ q(x, y) &= \sum_j \kappa_j q_{1j}(x) q_{2j}(y), \quad \kappa_j \geq 0, \quad \sum_j \kappa_j = 1, \end{aligned}$$

for $\mu_1 \times \mu_2$-almost-all pairs $(x, y) \in \mathscr{X} \times \mathscr{X}$. Here p_{1i} and q_{1j} are μ_1-densities for measures belonging to $\mathscr{P}_1$ and $\mathscr{Q}_1$ respectively (and analogously p_{2i} and q_{2j} are μ_2-densities for measures in $\mathscr{P}_2$ and $\mathscr{Q}_2$). This implies that we can write:

$$\begin{aligned} &\int p^{\alpha} q^{1-\alpha} \, d(\mu_1 \times \mu_2) \\ &\quad = \int \Bigl\{ \int \Bigl(\frac{\sum_i \lambda_i p_{1i}(x) p_{2i}(y)}{\sum_i \lambda_i p_{1i}(x)} \Bigr)^{\alpha} \Bigl(\frac{\sum_j \kappa_j q_{1j}(x) q_{2j}(y)}{\sum_j \kappa_j q_{1j}(x)} \Bigr)^{1-\alpha} d\mu_2(y) \Bigr\} \\ &\qquad \times \bigl(\sum_i \lambda_i p_{1i}(x) \bigr)^{\alpha} \bigl(\sum_j \kappa_j q_{1j}(x) \bigr)^{1-\alpha} d\mu_1(x), \end{aligned}$$

(where, as usual, the integrand of the inner integral is taken equal to zero whenever the μ_1-density equals zero). The inner integral is bounded by $\rho_\alpha(\mathscr{P}_2, \mathscr{Q}_2)$ for every fixed $x \in \mathscr{X}$. After substituting this upper bound the remaining integral is bounded by $\rho_\alpha(\mathscr{P}_1, \mathscr{Q}_1)$. □

Combining (3.39) with lemmas 3.14 and 3.13, we obtain the following theorem.

Theorem 3.9. *If P is a probability measure on $(\mathscr{X}, \mathscr{A})$ and $\mathscr{Q}$ is a dominated set of finite measures on $(\mathscr{X}, \mathscr{A})$, then for every $n \geq 1$ there exists a test $\phi_n : \mathscr{X}^n \to [0,1]$ such that for all $0 < \alpha < 1$:*

$$\sup_{Q \in \mathscr{Q}} \Big(P^n \phi_n + Q^n (1 - \phi_n) \Big) \leq \rho_\alpha(P, \mathscr{Q})^n.$$

The bound given by the theorem is useful only if $\rho_\alpha(P, \mathscr{Q}) < 1$. For probability measures P and Q we have

$$\rho_{1/2}(P, Q) = 1 - \tfrac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 \, d\mu,$$

and hence we might use the bound with $\alpha = 1/2$ if the Hellinger distance of $\mathrm{co}(\mathscr{Q})$ to P is positive. For a general finite measure Q the quantity $\rho_{1/2}(P, Q)$ may be bigger than 1 and depending on Q, the Hellinger transform $\rho_\alpha(P, Q)$ may even lie above 1 for every α. The following lemma shows that this is controlled by the (generalized) Kullback-Leibler divergence $-P \log(q/p)$.

Lemma 3.15. *For a probability measure P and a finite measure Q, the function $\rho : [0,1] \to \mathbb{R} : \alpha \mapsto \rho_\alpha(Q, P)$ is convex on $[0,1]$ with:*

$$\rho_\alpha(Q, P) \to P(q > 0), \quad \text{as } \alpha \downarrow 0, \qquad \rho_\alpha(Q, P) \to Q(p > 0), \quad \text{as } \alpha \uparrow 1,$$

and with the derivative at $\alpha = 0$ satisfying

$$\frac{d\rho_\alpha(Q, P)}{d\alpha} \Big|_{\alpha = 0} = P \log(q/p) 1_{q > 0},$$

(which may be equal to $-\infty$).

Proof The function $\alpha \mapsto e^{\alpha y}$ is convex on $(0,1)$ for all $y \in [-\infty, \infty)$, implying the convexity of $\alpha \mapsto \rho_\alpha(Q, P) = P(q/p)^\alpha$ on $(0,1)$. The function $\alpha \mapsto y^\alpha = e^{\alpha \log y}$ is continuous on $[0,1]$ for any $y > 0$, is decreasing for $y < 1$, increasing for $y > 1$ and constant for $y = 1$. By monotone convergence, as $\alpha \downarrow 0$,

$$Q\Big(\frac{p}{q}\Big)^\alpha 1_{\{0 < p < q\}} \uparrow Q\Big(\frac{p}{q}\Big)^0 1_{\{0 < p < q\}} = Q(0 < p < q).$$

By the dominated convergence theorem, with dominating function $(p/q)^{1/2} 1_{\{p \geq q\}}$ (which lies above $(p/q)^\alpha 1_{\{p \geq q\}}$ for $\alpha \leq 1/2$), we have (as $\alpha \to 0$):

$$Q\Big(\frac{p}{q}\Big)^\alpha 1_{\{p \geq q\}} \to Q\Big(\frac{p}{q}\Big)^0 1_{\{p \geq q\}} = Q(p \geq q).$$

Combining the two preceding displays above, we see that $\rho_{1-\alpha}(Q, P) = Q(p/q)^\alpha \to Q(p > 0)$ as $\alpha \downarrow 0$.

By the convexity of the function $\alpha \mapsto e^{\alpha y}$ the map $\alpha \mapsto f_\alpha(y) = (e^{\alpha y} - 1)/\alpha$ decreases as $\alpha \downarrow 0$, to $(d/d\alpha)|_{\alpha = 0} f_\alpha(y) = y$, for every y. For $y \leq 0$ we have $f_\alpha(y) \leq 0$, while for $y \geq 0$, by Taylor's formula,

$$f_\alpha(y) \leq \sup_{0 < \alpha' \leq \alpha} y e^{\alpha' y} \leq y e^{\alpha y} \leq \frac{1}{\epsilon} e^{(\alpha + \epsilon) y}.$$

Hence we conclude that $f_\alpha(y) \leq 0 \vee \epsilon^{-1} e^{(\alpha+\epsilon)y} 1_{y\geq 0}$. Consequently, we have:

$$\alpha^{-1}(e^{\alpha \log(q/p)} - 1) \downarrow \log(q/p), \qquad \text{as } \alpha \downarrow 0,$$

and is bounded above by $0 \vee \epsilon^{-1}(q/p)^{2\epsilon} 1_{q\geq p}$ for small $\alpha > 0$, which is P-integrable for $2\epsilon < 1$. We conclude that

$$\frac{1}{\alpha}(\rho_\alpha(Q,P) - \rho_0(Q,P)) = \frac{1}{\alpha} P\big((q/p)^\alpha - 1\big) 1_{q>0} \downarrow P \log(q/p) 1_{q>0},$$

as $\alpha \downarrow 0$, by the monotone convergence theorem. □

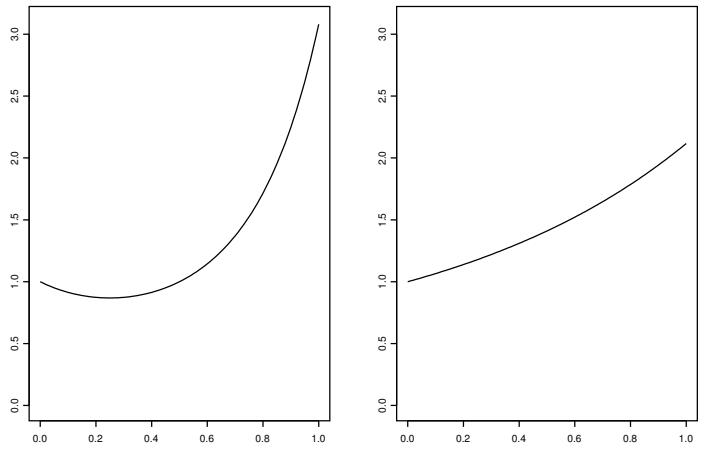

FIGURE 3.1 The Hellinger transforms $\alpha \mapsto \rho_\alpha(P,Q)$, for $P = N(0,2)$ and Q respectively the measure defined by $dQ = (dN(3/2,1)/dN(0,1))\, dP$ (left) and $dQ = (dN(3/2,1)/dN(1,1))\, dP$ (right). Intercepts with the vertical axis at the right and left of the graphs equal $P(q > 0)$ and $Q(p > 0)$ respectively. The slope at 0 equals (minus) the Kullback-Leibler divergence $P \log(p/q)$.

Two typical graphs of the Hellinger transform $\alpha \mapsto \rho_\alpha(Q,P)$ are shown in Figure 3.6.1 (corresponding to fitting a unit variance normal location model in a situation that the observations are sampled from an $N(0,2)$-distribution). For P a probability measure with $P \ll Q$ the Hellinger transform is equal to 1 at $\alpha = 0$, but will eventually increase to a level that is above 1 near $\alpha = 1$ if $Q(p > 0) > 1$. Unless the slope $P \log(p/q)$ is negative it will never decrease below the level 1. For probability measures P and Q this slope equals minus the Kullback-Leibler distance and hence is strictly negative unless $P = Q$. In that case the graph is strictly below 1 on $(0,1)$ and $\rho_{1/2}(P,Q)$ is a convenient choice to work with. For a general finite measure Q, the Hellinger transform $\rho_\alpha(Q,P)$ is guaranteed to assume values strictly less than 1 near $\alpha = 0$ provided that the Kullback-Leibler divergence $P \log(p/q)$ is negative, which is not automatically the case. For testing a composite alternative $\mathscr{Q}$, we shall need that this is the case uniformly in $Q \in \mathrm{co}(\mathscr{Q})$. For a convex alternative $\mathscr{Q}$ theorem 3.9 guarantees the existence of tests based on n observations with error probabilities bounded by $e^{-n\epsilon^2}$ if

$$\epsilon^2 \leq \sup_{0<\alpha<1} \sup_{Q\in\mathscr{Q}} \log \frac{1}{\rho_\alpha(Q,P)}.$$

In some of the examples we can achieve inequalities of this type by bounding the right side below by a (uniform) Taylor expansion of $\alpha \mapsto -\log \rho_\alpha(P,Q)$ in α near $\alpha = 0$. Such arguments are not mere technical generalizations: they can be necessary already to prove posterior consistency relative to misspecified standard parametric models.

If $P(q = 0) > 0$, then the Hellinger transform is strictly less than 1 at $\alpha = 0$ and hence good tests exist, even though it may be true that $\rho_{1/2}(P,Q) > 1$. The existence of good tests is obvious in this case, since we can reject Q if the observations land in the set $q = 0$.

In the above we have assumed that $\mathscr{Q}$ is dominated. If this is not the case, then the results go through, provided that we use Le Cam's generalized tests (see Le Cam (1986) [67]), *i.e.* we define

$$\pi(P,\mathscr{Q}) = \inf_{\phi} \sup_{Q\in\mathscr{Q}} \big(\phi P + (1-\phi)Q\big),$$

where the infimum is taken over the set of all continuous, positive linear maps $\phi : L_1(\mathscr{X},\mathscr{A}) \mapsto \mathbb{R}$ such that $\phi P \le 1$ for all probability measures P. This collection of functionals includes the linear maps that arise from integration of measurable functions $\phi : \mathscr{X} \mapsto [0,1]$, but may be larger. Such tests would be good enough for our purposes, but the generality appears to have little additional value for our application to misspecified models.

The next step is to extend the upper bound to alternatives $\mathscr{Q}$ that are possibly not convex. We are particularly interested in alternatives that are complements of balls around P in some metric. Let $L_1^+(\mathscr{X},\mathscr{A})$ be the set of finite measures on $(\mathscr{X},\mathscr{A})$ and let $\tau : L_1^+(\mathscr{X},\mathscr{A}) \times L_1^+(\mathscr{X},\mathscr{A}) \mapsto \mathbb{R}$ be such that $\tau(P,\cdot) : \mathscr{Q} \mapsto \mathbb{R}$ is a nonnegative function (written in a notation so as to suggest a distance from P to Q), which is dominated by $\sup_\alpha(-\log \rho_\alpha(P,\cdot))$ in the sense that for all $Q \in \mathscr{Q}$:

$$\tau^2(P,Q) \le \bar\tau^2(P,Q) := \sup_{0<\alpha<1} \log \frac{1}{\rho_\alpha(P,Q)}. \tag{3.40}$$

For $\epsilon > 0$ define $N_\tau(\epsilon,\mathscr{Q})$ to be the minimal number of convex subsets of $\{Q \in L_1^+(\mathscr{X},\mathscr{A}) : \bar\tau(P,Q) > \epsilon/2\}$ needed to cover $\{Q \in \mathscr{Q} : \epsilon < \tau(P,Q) < 2\epsilon\}$ and assume that $\mathscr{Q}$ is such that this number is finite for all $\epsilon > 0$. (The requirement that these convex subset have $\bar\tau$-distance $\epsilon/2$ to P is essential.) Then the following theorem applies.

Theorem 3.10. *Let P is a probability measure and $\mathscr{Q}$ is a dominated set of finite measures on $(\mathscr{X},\mathscr{A})$. Assume that $\tau : L_1^+(\mathscr{X},\mathscr{A}) \times L_1^+(\mathscr{X},\mathscr{A}) \mapsto \mathbb{R}$ satisfies (3.40). Then for all $\epsilon > 0$ and all $n \ge 1$, there exists a test ϕ_n such that for all $J \in \mathbb{N}$:*

$$\begin{aligned} P^n\phi_n &\le \sum_{j=1}^{\infty} N_\tau(j\epsilon,\mathscr{Q})\, e^{-nj^2\epsilon^2/4}, \\ \sup_{\{Q:\tau(P,Q)>J\epsilon\}} Q^n\big(1-\phi_n\big) &\le e^{-nJ^2\epsilon^2/4}. \end{aligned} \tag{3.41}$$

Proof Fix $n \ge 1$ and $\epsilon > 0$ and define $\mathscr{Q}_j = \{Q \in \mathscr{Q} : j\epsilon < \tau(P,Q) \le (j+1)\epsilon\}$. By assumption there exists, for every $j \ge 1$, a finite cover of $\mathscr{Q}_j$ by $N_j = N_\tau(j\epsilon,\mathscr{Q})$ convex sets

$C_{j,1}, \ldots, C_{j,N_j}$ of finite measures, with the further property that:

$$\inf_{Q \in C_{j,i}} \bar{\tau}(P,Q) > \frac{j\epsilon}{2}, \quad (1 \leq i \leq N_j).$$

According to theorem 3.9, for all $n \geq 1$ and for each set $C_{j,i}$, there exists a test $\phi_{n,j,i}$ such that for all $\alpha \in (0,1)$ we have:

$$P^n \phi_{n,j,i} \leq \rho_\alpha(P, C_{j,i})^n,$$
$$\sup_{Q \in C_{j,i}} Q^n\big(1 - \phi_{n,j,i}\big) \leq \rho_\alpha(P, C_{j,i})^n.$$

By (3.40), we have:

$$\sup_{Q \in C_{j,i}} \inf_{0<\alpha<1} \rho_\alpha(P,Q) = \sup_{Q \in C_{j,i}} e^{-\bar{\tau}^2(P,Q)} \leq e^{-j^2\epsilon^2/4}.$$

For fixed P and Q, the function $\alpha \mapsto \rho_\alpha(P,Q)$ is convex and can be extended continuously to a convex function on $[0,1]$. The function $Q \mapsto \rho_\alpha(P,Q)$ with domain $L_1^+(\mathscr{X}, \mathscr{A})$ is concave. By the minimax theorem (see *e.g.* Strasser (1985) [85], p. 239), the left side of the preceding display equals:

$$\inf_{0<\alpha<1} \sup_{Q \in C_{j,i}} \rho_\alpha(P,Q) = \inf_{0<\alpha<1} \rho_\alpha(P, C_{j,i}).$$

It follows that:

$$P^n \phi_{n,j,i} \vee \sup_{Q \in C_{j,i}} Q^n\big(1 - \phi_{n,j,i}\big) \leq e^{-nj^2\epsilon^2/4}.$$

Now define a new test function ϕ_n by:

$$\phi_n = \sup_{j \geq 1} \max_{1 \leq i \leq N_j} \phi_{n,j,i}.$$

Then, for every $J \geq 1$:

$$P^n \phi_n \leq \sum_{j=1}^{\infty} \sum_{i=1}^{N_j} P^n \phi_{n,j,i} \leq \sum_{j=1}^{\infty} N_j e^{-nj^2\epsilon^2/4},$$
$$\sup_{Q \in \mathscr{Q}} Q^n\big(1 - \phi_n\big) \leq \sup_{j \geq J} \max_{i \leq N_j} \sup_{Q \in C_{j,i}} Q^n\big(1 - \phi_{n,j,i}\big) \leq \sup_{j \geq J} e^{-nj^2\epsilon^2/4} = e^{-nJ^2\epsilon^2/4},$$

where $\mathscr{Q} = \{Q : \tau(P,Q) > J\epsilon\} = \cup_{j \geq J} \mathscr{Q}_j$. □

3.6.2 Application to misspecification

When applying the above in the proof for consistency in misspecified models the problem is to test the true distribution P_0 against measures $Q = Q(P)$ taking the form $dQ = (p_0/p^*)\, dP$ for $P \in \mathscr{P}$. In this case the Hellinger transform takes the form $\rho_\alpha(Q, P_0) = P_0(p/p^*)^\alpha$ and its right derivative at $\alpha = 0$ is equal to $P_0 \log(p/p^*)$. This is negative for every $P \in \mathscr{P}$ if and only if P^* is the point in $\mathscr{P}$ at minimal Kullback-Leibler divergence to P_0. This observation illustrates that the measure P^* in theorem 3.1 is necessarily a point of minimal Kullback-Leibler divergence, even if this is not explicitly assumed. We formalize this in the following lemma.

Lemma 3.16. *If P^* is such that $P_0 \log(p_0/p^*) < \infty$ and the right side of (3.11) is nonnegative, then $P_0 \log(p_0/p^*) \le P_0 \log(p_0/p)$ for every P with $P_0(p/p^*) < \infty$. Consequently, the covering numbers for testing $N_t(\epsilon, \mathscr{P}, d; P_0)$ can be finite only if P^* is a point of minimal Kullback-Leibler divergence relative to P_0.*

Proof The assumptions imply that $P_0(p^* > 0) = 1$. If $P_0(p = 0) > 0$, then $P_0 \log(p_0/p) = \infty$ and there is nothing to prove. Thus we may assume that p is also strictly positive under P_0. Then, in view of lemma 3.15, the function g defined by $g(\alpha) = P_0(p/p^*)^\alpha = \rho_\alpha(Q, P_0)$ is continuous on $[0, 1]$ with $g(0) = P_0(p > 0) = 1$ and the right side of (3.11) can be nonnegative only if $g(\alpha) \le 1$ for some $\alpha \in [0, 1]$. By convexity of g and the fact that $g(0) = 1$, this can happen only if the right derivative of g at zero is nonpositive. In view of lemma 3.15 this derivative is $g'(0+) = P_0 \log(p/p^*)$.

Finiteness of the covering numbers for testing for some $\epsilon > 0$ implies that the right side of (3.11) is nonnegative, as every $P \in \mathscr{P}$ must be contained in one of the sets B_i in the definition of $N_t(\epsilon, \mathscr{P}, d; P_0)$, in which case the right side of (3.11) is bounded below by $\epsilon^2/4$. □

If $P_0(p/p^*) \le 1$ for every $P \in \mathscr{P}$, then the measure Q defined by $dQ = (p_0/p^*)\, dP$ is a subprobability measure and hence by convexity the Hellinger transform $\alpha \mapsto \rho_\alpha(P_0, Q)$ is never above the level 1 and is strictly less than 1 at $\alpha = 1/2$ unless $P_0 = Q$. In such a case there appears to be no loss in generality to work with the choice $\alpha = 1/2$ only, leading to the distance d as in lemma 3.3. This lemma shows that this situation arises if $\mathscr{P}$ is convex.

The following theorem translates theorem 3.9 into the form needed for the proof of our main results. Recall the definition of the covering numbers for testing $N_t(\epsilon, \mathscr{P}, d; P_0)$ in (3.4).

Theorem 3.11. *Suppose $P^* \in \mathscr{P}$ and $P_0(p/p^*) < \infty$ for all $P \in \mathscr{P}$. Assume that there exists a nonincreasing function D such that for some $\epsilon_n \ge 0$ and every $\epsilon > \epsilon_n$:*

$$N_t(\epsilon, \mathscr{P}, d; P_0) \le D(\epsilon). \tag{3.42}$$

Then for every $\epsilon > \epsilon_n$ there exists a test ϕ_n (depending on $\epsilon > 0$) such that for every $J \in \mathbb{N}$,

$$\begin{aligned} P_0^n \phi_n &\le D(\epsilon) \frac{e^{-n\epsilon^2/4}}{1 - e^{-n\epsilon^2/4}}, \\ \sup_{\{P \in \mathscr{P} : d(P, P^*) > J\epsilon\}} Q(P)^n (1 - \phi_n) &\le e^{-nJ^2\epsilon^2/4}. \end{aligned} \tag{3.43}$$

Proof Define $\mathscr{Q}$ as the set of all finite measures $Q(P)$ as P ranges over $\mathscr{P}$ (where $p_0/p^* = 0$ if $p_0 = 0$) and define $\tau(Q_1, Q_2) = d(P_1, P_2)$. Then $Q(P^*) = P_0$ and hence $d(P, P^*) = \tau(Q(P), P_0)$. Identify P of theorem 3.9 with the present measure P_0. By the definitions (3.4) and (3.40) $N_\tau(\epsilon, \mathscr{Q}) \le N_t(\epsilon, \mathscr{P}, d) \le D(\epsilon)$ for every $\epsilon > \epsilon_n$. Therefore, the test function guaranteed to exist by theorem 3.9 satisfies:

$$P_0^n \phi_n \le \sum_{j=1}^{\infty} D(j\epsilon)\, e^{-nj^2\epsilon^2/4} \le D(\epsilon) \sum_{j=1}^{\infty} e^{-nj^2\epsilon^2/4},$$

because D is nonincreasing. This can be bounded further (as in the assertion) since for all $0 < x < 1$, $\sum_{n\geq 1} x^{n^2} \leq x/(1-x)$. The second line in the assertion is simply the second line in (3.41). □

3.7 Proofs of the main theorems

The following lemma is analogous to lemma 8.1 in Ghosal *et al.* (2000) [39] and can be proved in the same manner.

Lemma 3.17. *For given $\epsilon > 0$ and $P_0 \in \mathscr{P}$ define $B(\epsilon)$ by (3.5). Then for every $C > 0$ and probability measure Π on $\mathscr{P}$:*

$$P_0^n\Big(\int \prod_{i=1}^n \frac{p}{p^*}(X_i)\, d\Pi(P) < \Pi(B(\epsilon, P^*; P_0))e^{-n\epsilon^2(1+C)}\Big) \leq \frac{1}{C^2 n\epsilon^2}.$$

Proof of theorem 3.2 In view of (3.7), the conditions of theorem 3.11 are satisfied, with the function $D(\epsilon) = e^{n\epsilon_n^2}$, (*i.e.* constant in $\epsilon > \epsilon_n$). Let ϕ_n be the test as in the assertion of this theorem for $\epsilon = M\epsilon_n$ and M a large constant, to be determined later in the proof.

For $C > 0$, also to be determined later in the proof, let Ω_n be the event

$$\int_{\mathscr{P}} \prod_{i=1}^n \frac{p}{p^*}(X_i)\, d\Pi(P) \geq e^{-(1+C)n\epsilon_n^2}\Pi\big(B(\epsilon_n, P^*; P_0)\big). \tag{3.44}$$

Then $P_0^n(\mathscr{X}^n \setminus \Omega_n) \leq 1/(C^2 n\epsilon_n^2)$, by lemma 3.17.

Set $\hat{\Pi}_n(\epsilon) = \Pi_n\big(P \in \mathscr{P} : d(P, P^*) > \epsilon \mid X_1, \ldots, X_n\big)$. For every $n \geq 1$ and $J \in \mathbb{N}$ we can decompose:

$$\begin{aligned} P_0^n\hat{\Pi}_n(JM\epsilon_n) &= P_0^n\Big(\hat{\Pi}_n(JM\epsilon_n)\phi_n\Big) \\ &\quad + P_0^n\Big(\hat{\Pi}_n(JM\epsilon_n)(1-\phi_n)1_{\Omega_n^c}\Big) + P_0^n\Big(\hat{\Pi}_n(JM\epsilon_n)(1-\phi_n)1_{\Omega_n}\Big). \end{aligned} \tag{3.45}$$

We estimate the three terms on the right side separately. Because $\hat{\Pi}_n(\epsilon) \leq 1$, the middle term is bounded by $1/(C^2 n\epsilon_n^2)$. This converges to zero as $n\epsilon_n^2 \to \infty$ for fixed C and/or can be made arbitrarily small by choosing a large constant C if $n\epsilon_n^2$ is bounded away from zero.

By the first inequality in (3.43), the first term on the right of (3.45) is bounded by:

$$P_0^n\Big(\hat{\Pi}_n(JM\epsilon_n)\phi_n\Big) \leq P_0^n\phi_n \leq \frac{e^{(1-M^2/4)n\epsilon_n^2}}{1 - e^{-nM^2\epsilon_n^2/4}}.$$

For sufficiently large M, the expression on the right-hand side is bounded above by $2e^{-n\epsilon_n^2 M^2/8}$ for sufficiently large n and hence can be made arbitrarily small by choice of M, or converges to 0 for fixed M if $n\epsilon_n^2 \to \infty$.

Estimation of the third term on the right of (3.45) is more involved. Because $P_0(p^* > 0) = 1$, we can write

$$\begin{aligned} &P_0^n\Big(\hat{\Pi}_n(JM\epsilon_n)(1-\phi_n)1_{\Omega_n}\Big) \\ &\qquad = P_0^n(1-\phi_n)1_{\Omega_n}\left[\frac{\int_{d(P,P^*)>JM\epsilon_n} \prod_{i=1}^n (p/p^*)(X_i)\, d\Pi(P)}{\int_{\mathscr{P}} \prod_{i=1}^n (p/p^*)(X_i)\, d\Pi(P)}\right], \end{aligned} \tag{3.46}$$

where we have written the arguments X_i for clarity. By the definition of Ω_n the integral in the denominator is bounded below by $e^{-(1+C)n\epsilon_n^2}\Pi(B(\epsilon_n, P^*; P_0))$. Inserting this bound, writing $Q(P)$ for the measure defined by $dQ(P) = (p_0/p^*)\,dP$, and using Fubini's theorem, we can bound the right side of the preceding display by

$$\frac{e^{(1+C)n\epsilon_n^2}}{\Pi\big(B(\epsilon_n, P^*; P_0)\big)} \int_{d(P,P^*)>JM\epsilon_n} Q(P)^n(1-\phi_n)\,d\Pi(P). \tag{3.47}$$

Setting $\mathscr{P}_{n,j} = \{P \in \mathscr{P} : M\epsilon_n j < d(P, P^*) \le M\epsilon_n(j+1)\}$, we can decompose $\{P : d(P, P^*) > JM\epsilon_n\} = \cup_{j\ge J}\mathscr{P}_{n,j}$. The tests ϕ_n have been chosen to satisfy the inequality $P^n\big(L_n(1-\phi_n)\big) \le e^{-nj^2M^2\epsilon_n^2/4}$ uniformly in $P \in \mathscr{P}_{n,j}$. (*c.f.* the second inequality in (3.43).) It follows that the preceding display is bounded by

$$\frac{e^{(1+C)n\epsilon_n^2}}{\Pi\big(B(\epsilon_n, P^*; P_0)\big)} \sum_{j\ge J} e^{-nj^2M^2\epsilon_n^2/4}\Pi_n(\mathscr{P}_{n,j}) \le \sum_{j\ge J} e^{(1+C)n\epsilon_n^2 + n\epsilon_n^2M^2j^2/8 - nj^2M^2\epsilon_n^2/4},$$

by (3.13). For fixed C and sufficiently large M this converges to zero if $n\epsilon_n^2$ is bounded away from zero and $J = J_n \to \infty$. □

Proof of theorem 3.1 Because Π is a probability measure, the numerator in (3.13) is bounded above by 1. Therefore, the prior mass condition (3.13) is implied by the prior mass condition (3.6). We conclude that the assertion of theorem 3.1, but with $M = M_n \to \infty$, follows from theorem 3.2. That in fact it suffices that M is sufficiently large follows by inspection of the preceding proof. □

Proof of theorem 3.4 The proof of this theorem follows the same steps as the preceding proofs. A difference is that we cannot appeal to the preparatory lemmas and theorems to split the proof in separate steps. The shells $\mathscr{P}_{n,j} = \{P \in P : Mj\epsilon_n < d(P, \mathscr{P}^*) < M(j+1)\epsilon_n\}$ must be covered by sets $B_{n,j,i}$ as in the definition (3.18) and for each such set we use the appropriate element $P^*_{n,j,i} \in \mathscr{P}^*$ to define a test $\phi_{n,j,i}$ and to rewrite the left side of (3.46). We omit the details. □

Lemma 3.18 is used to upper bound the Kullback Leibler divergence and the expectation of the squared logarithm by a function of the L_1-norm. A similar lemma was presented in Wong and Shen (1995) [97], where both p and q were assumed to be densities of probability distributions. We generalise this result to the case where q is a finite measure and we are forced to use the L_1 instead of the Hellinger distance.

Lemma 3.18. *For every $b > 0$ there exists a constant $\epsilon_b > 0$ such that for every probability measure P and finite measure Q with $0 < h^2(p,q) < \epsilon_b P(p/q)^b$,*

$$P\log\frac{p}{q} \lesssim h^2(p,q)\Big(1 + \frac{1}{b}\log_+\frac{1}{h(p,q)} + \frac{1}{b}\log_+ P\Big(\frac{p}{q}\Big)^b\Big) + \|p-q\|_1,$$
$$P\Big(\log\frac{p}{q}\Big)^2 \lesssim h^2(p,q)\Big(1 + \frac{1}{b}\log_+\frac{1}{h(p,q)} + \frac{1}{b}\log_+ P\Big(\frac{p}{q}\Big)^b\Big)^2.$$

Proof The function $r : (0, \infty) \to \mathbb{R}$ defined implicitly by $\log x = 2(\sqrt{x} - 1) - r(x)(\sqrt{x} - 1)^2$ possesses the following properties:

- r is nonnegative and decreasing.
- $r(x) \sim \log(1/x)$ as $x \downarrow 0$, whence there exists $\epsilon' > 0$ such that $r(x) \leq 2\log(1/x)$ on $[0, \epsilon']$. (A computer graph indicates that $\epsilon' = 0.4$ will do.)
- For every $b > 0$ there exists $\epsilon''_b > 0$ such that $x^b r(x)$ is increasing on $[0, \epsilon''_b]$. (For $b \geq 1$ we may take $\epsilon''_b = 1$, but for b close to zero ϵ''_b must be very small.)

In view of the definition of r and the first property, we can write

$$\begin{aligned} P\log\frac{p}{q} &= -2P\Big(\sqrt{\frac{q}{p}} - 1\Big) + Pr\Big(\frac{q}{p}\Big)\Big(\sqrt{\frac{q}{p}} - 1\Big)^2 \\ &\leq h^2(p,q) + 1 - \int q\,d\mu + Pr\Big(\frac{q}{p}\Big)\Big(\sqrt{\frac{q}{p}} - 1\Big)^2 \\ &\leq h^2(p,q) + \|p - q\|_1 + r(\epsilon)h^2(p,q) + Pr\Big(\frac{q}{p}\Big)1\Big\{\frac{q}{p} \leq \epsilon\Big\}, \end{aligned}$$

for any $0 < \epsilon \leq 4$, where we use that $|\sqrt{q/p} - 1| \leq 1$ if $q/p \leq 4$. Next we choose $\epsilon \leq \epsilon''_b$ and use the third property to bound the last term on the right by $P(p/q)^b \epsilon^b r(\epsilon)$. Combining the resulting bound with the second property we then obtain, for $\epsilon \leq \epsilon' \wedge \epsilon''_b \wedge 4$,

$$P\log\frac{p}{q} \leq h^2(p,q) + \|p - q\|_1 + 2\log\frac{1}{\epsilon}h^2(p,q) + 2\epsilon^b \log\frac{1}{\epsilon}\,P\Big(\frac{p}{q}\Big)^b.$$

For $\epsilon^b = h^2(p,q)/P(p/q)^b$ the second and third terms on the right take the same form. If $h^2(p,q) < \epsilon_b P(p/q)^b$ for a sufficiently small ϵ_b, then this choice is eligible and the first inequality of the lemma follows. Specifically, we can choose $\epsilon_b \leq (\epsilon' \wedge \epsilon''_b \wedge 4)^b$.

To prove the second inequality we first note that, since $|\log x| \leq 2|\sqrt{x} - 1|$ for $x \geq 1$,

$$P\Big(\log\frac{p}{q}\Big)^2 1\Big\{\frac{q}{p} \geq 1\Big\} \leq 4P\Big(\sqrt{\frac{q}{p}} - 1\Big)^2 = 4h^2(p,q).$$

Next, with r as in the first part of the proof,

$$\begin{aligned} P\Big(\log\frac{p}{q}\Big)^2 1\Big\{\frac{q}{p} \leq 1\Big\} &\leq 8P\Big(\sqrt{\frac{q}{p}} - 1\Big)^2 + 2Pr^2\Big(\frac{q}{p}\Big)\Big(\sqrt{\frac{q}{p}} - 1\Big)^4 1\Big\{\frac{q}{p} \leq 1\Big\} \\ &\leq 8h^2(p,q) + 2r^2(\epsilon)h^2(p,q) + 2\epsilon^b r^2(\epsilon)P\Big(\frac{p}{q}\Big)^b, \end{aligned}$$

for $\epsilon \leq \epsilon''_{b/2}$, in view of the third property of r. (The power of 4 in the first line of the array can be lowered to 2 or 0, as $|\sqrt{q/p} - 1| \leq 1$.) We can use the second property of r to bound $r(\epsilon)$ and next choose $\epsilon^b = h^2(p,q)/P(p/q)^b$ to finish the proof. Specifically, we can choose $\epsilon_b \leq (\epsilon' \wedge \epsilon''_{b/2})^b$. □

Chapter 4

Errors-In-Variables regression

In the first five sections of this chapter, we consider the asymptotic behaviour of the posterior distribution for the errors-in-variables model. The model describes measurements consisting of paired observations (X, Y) that are represented in terms of an unobserved Z. The random variable Z is related to X directly and to Y through a regression function f, both perturbed by Gaussian errors. We assume that Z falls into a (known) bounded subset of the real line with probability one, but otherwise leave its distribution unconstrained. In the semi-parametric literature, the regression function comes from a parametric (see Taupin (2001) [86]), or even linear (see, *e.g.* Anderson (1984) [2]) class of functions. In the following, we broaden that assumption to non-parametric regression classes, discussing the errors-in-variables problem also for Lipschitz and smooth functions.

Hence, the formulation we use involves two non-parametric components, the distribution of Z and the regression function f. We give Hellinger rates of convergence for the posterior distribution of the errors-in-variables density in non-parametric and parametric regression classes, using the posterior rate-of-convergence theorem 1.8 (or rather, a version based on the Hellinger metric entropy, *c.f.* Ghosal *et al.* (2000) [39]). Conditions that bound the rate of convergence can be decomposed into two terms, one for each of the non-parametric components of the model. The rate is then determined by the term that dominates the bound. A version of the first five sections of this chapter is to be submitted to the *Annals of Statistics* for publication.

Even in the case of a parametric class of regression functions, the rate is $1/\sqrt{n}$ only up to a logarithmic correction, due to the (non-parametric) distribution of Z. Nevertheless, in the semi-parametric analyses referred to above, point-estimation of parametric regression functions proceeds (efficiently) at rate $1/\sqrt{n}$. In the last section of this chapter, we use the main result of chapter 2, theorem 2.1, to discuss possibilities for the derivation of analogous results in a Bayesian setting. The strategy that is outlined follows the least-favourable approach that is central in the semi-parametric literature and gives indications for the way to a semi-parametric Bernstein-Von-Mises theorem.

A Bayesian analysis of Errors-In-Variables regression

B.J.K. KLEIJN AND A.W. VAN DER VAART

UC Berkeley Statistics Department
Free University Amsterdam

Abstract

We consider the asymptotic behaviour of the posterior distribution for the structural errors-in-variables model with non-parametric and parametric regression classes. Generically, the formulation involves two non-parametric components, one being the distribution of the unobserved random variable and the other the regression function f. The prior on the former derives from a Dirichlet process and the prior on the latter is a so-called net prior. Entropy and prior-mass conditions that bound the rate are decomposed into two terms, one for each of the non-parametric components. The rate at which the posterior for the errors-in-variables density converges in Hellinger distance is then determined by the term that dominates the bounds.

4.1 Introduction

The errors-in-variables model is intended for the study of samples consisting of paired observations (X, Y), assumed to be distributed as follows:

$$\begin{aligned} X &= Z + e_1, \\ Y &= f(Z) + e_2, \end{aligned} \tag{4.1}$$

where (e_1, e_2) and Z are independent and f belongs to a family of regression functions. Usually, the distribution of the errors (e_1, e_2) is assumed to be known up to a (finite-dimensional) parameter σ whereas the distribution F of Z is completely unknown in the most general case. The primary interest lies in estimation of the regression function f from a *i.i.d.* sample of pairs (X_1, Y_1), (X_2, Y_2), $\ldots, (X_n, Y_n)$ in the presence of the nuisance parameter F. Applications

include all situations in which a functional dependence between measurements with errors is to be established. A prototypical example of a situation in which the errors-in-variables model applies, arises in experimental physics: often, the aim of an experiment is to establish a functional dependence between two physical quantities rather than to infer from their distribution directly. The experiment involves repeated, simultaneous measurement of the two quantities of interest, with errors that are due to the measurement apparatus. With regard to the distribution of the latter, measurement of prepared, known signals or background noise provide detailed information.

The primary difference between errors-in-variables and ordinary regression using a set of design points $x_1, \ldots, x_n$, is the stochastic nature of the variable X. Regarding X, the variable e_1 is referred to as the "random error", whereas the variability of Z is said to be the "systematic error" (Anderson (1984) [2]). Kendall and Stuart (1979) [54] distinghuish between the "functional" errors-in-variables problem, in which Z is non-stochastic, taking on the values of 'design points' $z_1, \ldots, z_n$, and the "structural" errors-in-variables problem, in which Z is stochastic. Best known is linear errors-in-variables regression, in which f is assumed to depend linearly on z (see, *e.g.* [2] for an extensive overview of the literature). Efficient estimators for the parameters of f have been constructed by Bickel and Ritov (1987) [13], Bickel *et al.* (1998) [14] and Van der Vaart (1988, 1996) [87, 88]. Errors-in-variables regression involving a parametric family of non-linear regression functions has been analysed by Taupin and others (see Taupin (2001) [86] and references therein). In Fan and Troung (1993) [31], the rate of convergence (in a weighted L_2-sense) of Nadaraya-Watson-type kernel estimators for the conditional expectation of Y given Z (and hence for the regression function) are considered using deconvolution methods.

In this paper we analyse the structural errors-in-variables problem for non-parametric families of regression functions in a Bayesian setting; we consider the behaviour of posterior distributions for the parameter (σ, f, F) in the asymptotic limit. It is stressed that in this formulation, the errors-in-variables problem has two non-parametric components, one being the distribution of the underlying variable Z and the other the regression function. The emphasis lies on the interplay between these two non-parametric aspects of the model, as illustrated by their respective contributions to the rate of convergence (see, *e.g.* theorems 4.3 and 4.4).

4.1.1 Model definition

We assume throughout this paper that there is some known constant $A > 0$ such that $Z \in [-A, A]$ with probability one. Furthermore, we assume (unless indicated otherwise) that the errors e_1 and e_2 are independent and distributed according to the same normal distribution Φ_σ on $\mathbb{R}$ with mean zero and variance σ^2 (*i.e.* a special case of *restricted Gaussian errors* in the terminology of [13]). Writing φ_σ for the normal density of both e_1 and e_2, the model consists of a family of distributions for the observations (X, Y), parametrized by $(\sigma, f, F) \in I \times \mathscr{F} \times D$,

where it is assumed that:

(a) I is a closed interval in the positive reals, bounded away from zero and infinity, *i.e.* $I = [\underline{\sigma}, \overline{\sigma}] \subset (0, \infty)$.

(b) D is the collection of all probability distributions on the compact symmetric interval $[-A, A]$, parametrized by all corresponding Stieltjes functions F.

(c) $\mathscr{F} \subset C_B[-A, A] \subset C[-A, A]$ is a bounded family of continuous regression functions $f : [-A, A] \to [-B, B]$. We shall distinguish various cases by further requirements, including equicontinuity, Lipschitz- and smoothness-bounds. Also considered is the parametric case, in which $\mathscr{F}$ is parametrized by a subset of $\mathbb{R}^k$.

For all $(\sigma, f, F) \in I \times \mathscr{F} \times D$, we define the following convoluted density for the distribution of observated pair (X, Y):

$$p_{\sigma,f,F}(x, y) = \int_{\mathbb{R}} \varphi_\sigma\big(x - z\big)\varphi_\sigma\big(y - f(z)\big)\, dF(z), \tag{4.2}$$

for all $(x, y) \in \mathbb{R}^2$.

It is stressed that when we speak of the errors-in-variables *model* $\mathscr{P}$, we refer to the collection of probability measures $P_{\sigma,f,F}$ on $\mathbb{R}^2$ defined by the Lebesgue-densities parametrized in the above display (rather than the parameter space $I \times \mathscr{F} \times D$). In many cases we regard $\mathscr{P}$ as a metric space, using either the Hellinger metric or $L_1(\mu)$-norm. As far as the parameter space is concerned, the model may not be identifiable: if, for given $F \in D$, two regression functions $f, g \in \mathscr{F}$ differ only on a set of F-measure zero, the corresponding densities $p_{\sigma,f,F}$ and $p_{\sigma,g,F}$ are equal on all of $\mathbb{R}^2$ (for all $\sigma \in I$). Determination of the true regression function f_0 based on an *i.i.d.* P_0-distributed sample can therefore be done only F_0-almost-everywhere (where $P_0 = P_{\sigma_0,f_0,F_0}$). Ultimately, the results we give are based on the Hellinger distance, which, in the present circumstances, gives rise to a semi-metric on the parameter space $I \times \mathscr{F} \times D$ for the same reason. The 'well-known' identifiability problems in the linear errors-in-variables model (see *e.g.* Reiersøl (1950) [80]) arising due to interchangability of Gaussian components of the distribution of Z with the error-distribution (see also [2] and [13]) do not occur in our considerations, because we assume the distribution of Z to be compactly supported.

4.1.2 Bayesian rates of convergence

Conditions for the theorem on Bayesian rates of convergence that is used in this paper are formulated in terms of a specific kind of Kullback-Leibler neighbourhoods of the true distribution $P_0 \in \mathscr{P}$ and Hellinger covering numbers for the model. For all $\epsilon > 0$ we define the following neighbourhoods:

$$B(\epsilon; P_0) = \Big\{P \in \mathscr{P} : -P_0 \log \frac{p}{p_0} \le \epsilon^2,\ P_0\Big(\log \frac{p}{p_0}\Big)^2 \le \epsilon^2\Big\}. \tag{4.3}$$

Denote by $N(\epsilon, \mathscr{P}, H)$ the covering numbers with respect to the Hellinger metric on $\mathscr{P}$, *i.e.* the minimal number of Hellinger balls of radius $\epsilon > 0$ needed to cover the model $\mathscr{P}$.

Theorem 4.1. *Let $\mathscr{P}$ be a model and assume that the sample $U_1, U_2, \ldots$ is i.i.d. P_0-distributed for some $P_0 \in \mathscr{P}$. For a given prior Π, suppose that there exists a sequence of strictly positive numbers ϵ_n with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ and constants $R_1, R_2 > 0$, such that:*

$$\Pi\big(B(\epsilon_n; P_0)\big) \geq e^{-R_1 n\epsilon_n^2}, \tag{4.4}$$

$$\log N\big(\epsilon_n, \mathscr{P}, H\big) \leq R_2 n\epsilon_n^2, \tag{4.5}$$

for all large enough n. Then, for every sufficiently large constant M, the posterior distribution satisfies:

$$\Pi_n\big(P \in \mathscr{P} : H(P, P_0) \geq M\epsilon_n \,\big|\, U_1, \ldots, U_n \big) \to 0, \tag{4.6}$$

as $n \to \infty$, in P_0-expectation.

The proof of this theorem can be found in Ghosal, Ghosh and Van der Vaart (2000) [39]. The two main conditions of theorem 4.1 are a prior-mass condition (4.4) and an entropy condition (4.5). Below we discuss the background of these conditions in turn. The lower bound (4.4) on the prior mass of $B(\epsilon; P_0)$ requires that the prior measure assigns a certain minimal share of its total mass to the Kullback-Leibler neighbourhoods of P_0 defined above. Since P_0 is unknown, a demonstration that (4.4) is satisfied usually requires proof that it is satisfied for all P in the model, *i.e.* the prior in question places enough mass in neighbourhoods of *all* points in the model. Therefore, a rough understanding of (4.4) for the best achievable rate ϵ_n is that a corresponding prior spreads its mass 'uniformly' over $\mathscr{P}$. The purpose of the entropy condition is to measure the complexity of the model, with faster-growing entropies leading to slower rates of convergence. From a more technical perspective, the entropy condition guarantees the existence of sufficiently powerful test functions to separate P_0 from complements of Hellinger neighbourhoods. The minimal ϵ_n satisfying $\log N(\epsilon_n, \mathscr{P}, H) \leq n\epsilon_n^2$ is roughly the fastest rate of convergence for estimating a density in the model $\mathscr{P}$ relative to the Hellinger distance obtainable by any method of estimation (*c.f.* Birgé (1983) [15]). Based on the sequence of posterior distributions in (4.6), point-estimator sequences converging at the same rate can be obtained (see, *e.g.* theorem 2.5 in [39]).

The assumption that the model is well-specified, *i.e.* $P_0 \in \mathscr{P}$, can be relaxed. In Kleijn and Van der Vaart (2003) [57], the above theorem is given in the case of a misspecified model. We do not give misspecified versions of the results, although we believe that the conditions of the necessary theorems in [57] are met in the model we consider.

Notation and conventions

We denote the Lebesgue measure on $\mathbb{R}^2$ by μ; for a probability distribution P on $\mathbb{R}^2$ dominated by μ, the corresponding density is denoted p. The Hellinger distance $H(P, Q)$ between two measures dominated by μ is defined as the $L_2(\mu)$-distance between the square-roots of

their densities, *i.e.* without the additional factors favoured by some authors. The space of continuous, real-valued functions on the interval $[-A, A] \subset \mathbb{R}$ is denoted $C[-A, A]$; the class of functions in $C[-A, A]$ that is uniformly bounded by a constant $B > 0$ is denoted $C_B[-A, A]$. When considering families of regression functions, $\|\cdot\|$ denotes the uniform norm over the interval $[-A, A]$. The space of all real-valued polynomials of degree n on $[-A, A]$ is denoted P_n. The Euclidean norm on $\mathbb{R}^k$ is denoted by $\|\cdot\|_{\mathbb{R}^k}$. The k-th derivative of a suitably differentiable function $f : [-A, A] \to \mathbb{R}$ is denoted $f^{(k)}$.

4.2 Main results

We consider regression classes $\mathscr{F} \subset C_B[-A, A]$, *i.e.* there exists a (known) constant $B > 0$ such that $\|f\| \leq B$ for all $f \in \mathscr{F}$, with a constraint that guarantees equicontinuity and allows for the establishment of bounds on covering numbers with respect to the uniform norm. We distinguish several non-parametric and parametric examples of such classes below, but remark that other regression classes for which bounds on covering numbers exist, can also be used.

(i) $\mathrm{Lip}_M(\alpha)$ (for some $M > 0$ and $0 < \alpha \leq 1$), the class of all Lipschitz functions $f \in C_B[-A, A]$ with constant M and exponent α, *i.e.*

$$\left|f(z) - f(z')\right| \leq M|z - z'|^\alpha, \tag{4.7}$$

for all $z, z' \in [-A, A]$.

(ii) $D_{\alpha,M}(q)$ (for some $0 < \alpha \leq 1$, $M > 0$ and an integer $q \geq 1$), the class of all q times differentiable functions $f \in C_B[-A, A]$ for which the q-th derivative $f^{(q)}$ belongs to $\mathrm{Lip}_M(\alpha)$.

(iii) $\mathscr{F}_\Theta$, a parametric class of regression functions which forms a subset of $\mathrm{Lip}_M(\alpha)$ for some $\alpha \in (0, 1]$ and $M > 0$. We assume that there exists a bounded, open subset $\Theta \subset \mathbb{R}^k$ for some $k \geq 1$ such that $\mathscr{F}_\Theta = \{f_\theta : \theta \in \Theta\}$. Furthermore, we assume that the map $\theta \mapsto f_\theta$ is Lipschitz-continuous, *i.e.* there exist constants $L > 0$ and $\rho \in (0, 1]$ such that for all $\theta_1, \theta_2 \in \Theta$:

$$\|f_{\theta_1} - f_{\theta_2}\| \leq L\|\theta_1 - \theta_2\|_{\mathbb{R}^k}^\rho. \tag{4.8}$$

Often, it is more convenient to unify cases (i) and (ii) above, by considering the family of classes $C_{\beta,L}[-A, A]$ defined as follows. For given $\beta > 0$ and $L > 0$, we define $\underline{\beta}$ to be the greatest integer such that $\underline{\beta} < \beta$ and we consider, for suitable functions $f : [-A, A] \to \mathbb{R}$, the norm:

$$\|f\|_\beta = \max_{k \leq \underline{\beta}} \|f^{(k)}\| + \sup_{z_1, z_2} \frac{\left|f^{(\underline{\beta})}(z_1) - f^{(\underline{\beta})}(z_2)\right|}{|z_1 - z_2|^{\beta - \underline{\beta}}},$$

where the supremum is taken over all pairs $(z_1, z_2) \in [-A, A]^2$ such that $z_1 \neq z_2$. The class $C_{\beta,L}[-A, A]$ is then taken to be the collection of all continuous $f : [-A, A] \to \mathbb{R}$ for which $\|f\|_\beta \leq L$. Note that for $0 < \beta \leq 1$, $\underline{\beta} = 0$ and $C_{\beta,L}[-A, A]$ is a Lipschitz class bounded by L;

if $\beta > 1$, differentiability of a certain order is implied, as well as boundedness of all derivatives and a Lipschitz property for the highest derivative.

As indicated in subsection 4.1.2, the Hellinger rate of convergence ϵ_n is bounded by two conditions, one related to the small-ϵ behaviour of the (Hellinger) entropy of the model, the other by the small-ϵ behaviour of the prior mass in Kullback-Leibler neighbourhoods of the form (4.3). The first condition is considered in section 4.3: theorem 4.3 says that the Hellinger covering number of the errors-in-variables model has an upper bound that consists of two terms, one resulting from the (σ, F)-part of the model and the other from minimal covering of the regression class:

$$\log N(\epsilon, \mathscr{P}, H) \leq L_0 \Big(\log \frac{1}{\epsilon} \Big)^3 + \log N(L\epsilon, \mathscr{F}, \| \, . \, \|), \tag{4.9}$$

for small $\epsilon > 0$ and some constants $L, L_0 > 0$. If the regression class $\mathscr{F}$ is 'small' enough, in the sense that the first term in the entropy bound displayed above dominates in the limit $\epsilon \to 0$, the candidate rates of convergence ϵ_n are parametric up to a logarithmic factor.

Lemma 4.1. *If there exists a constant $L_1 > 0$ such that:*

$$\log N(\epsilon, \mathscr{F}, \| \, . \, \|) \leq L_1 \Big(\log \frac{1}{\epsilon} \Big)^3, \tag{4.10}$$

for small enough $\epsilon > 0$, then the entropy condition (4.5) is satisfied by the sequence:

$$\epsilon_n = n^{-1/2} (\log n)^{3/2}, \tag{4.11}$$

for large enough n.

Proof Under the above assumption, $\log N(\epsilon, \mathscr{P}, H)$ is upper bounded by the first term in (4.9) with a larger choice for the constant. Note that the sequence ϵ_n as defined in (4.11) satisfies $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$. Also note that $\epsilon_n \geq 1/n$ for large enough n, so that for some $L > 0$,

$$\log N(\epsilon_n, \mathscr{F}, \| \, . \, \|) \leq \log N(1/n, \mathscr{F}, \| \, . \, \|) \leq L (\log n)^3,$$

and $n\epsilon_n^2 = (\log n)^3$, which proves that ϵ_n satisfies (4.5). □

It is also possible that the small-ϵ behaviour of the errors-in-variables entropy is dominated by the covering numbers of the regression class. In that case the *r.h.s.* of (4.9) is replaced by a single term proportional to $\log N(L\epsilon, \mathscr{F}, \| \, . \, \|)$ for small enough ϵ. If the regression functions constitute a Lipschitz or smoothness class, lemma 4.13 gives the appropriate upper bound for the entropy, leading to the following candidate rates of convergence.

Lemma 4.2. *For an errors-in-variables model $\mathscr{P}$ based on a regression class $C_{\beta,M}[-A, A]$, the entropy condition (4.5) is satisfied by the sequence:*

$$\epsilon_n = n^{-\frac{\beta}{2\beta+1}}, \tag{4.12}$$

for large enough n.

Proof As argued above, the Hellinger entropy of the errors-in-variables model is upper-bounded as follows:

$$\log N(\epsilon, \mathscr{P}, H)) \leq \frac{K}{\epsilon^{1/\beta}},$$

for some constant $K > 0$ and small enough ϵ. The sequence ϵ_n satisfies $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$. Furthermore, note that:

$$\log N(\epsilon_n, \mathscr{P}, H) \leq K\, n^{1/(2\beta+1)} = Kn \cdot n^{-\frac{2\beta}{2\beta+1}} = Kn\epsilon_n^2,$$

for large enough n. □

Similar reasoning applies to condition (4.4) for the small-ϵ behaviour of the prior mass of Kullback-Leibler neighbourhoods of the form (4.3). Section 4.4 discusses the necessary lemmas in detail. We define priors Π_I, $\Pi_{\mathscr{F}}$ and Π_D on the parametrizing spaces I, $\mathscr{F}$ and D respectively and choose the prior Π on the model $\mathscr{P}$ as induced by their product under the map $(\sigma, f, F) \mapsto P_{\sigma,f,F}$ (which is measurable, as shown in lemma 4.10). The prior Π_I is chosen as a probability measure on I with continuous and strictly positive density with respect to the Lebesgue measure on I. Priors for the various regression classes discussed in the beginning of this section are discussed in subsection 4.5.2. The prior Π_D on D is based on a Dirichlet process with base measure α which has a continuous and strictly positive density on all of $[-A, A]$.

As with the covering numbers discussed above, we find (see theorem 4.4) that (the logarithm of) the prior mass of Kullback-Leibler neighbourhoods is lower bounded by two terms, one originating from the prior on the regression class and the other from the priors on the remaining parameters in the model:

$$\log \Pi\Big(B\big(K\delta \log(1/\delta); P_0\big)\Big) \geq -c\Big(\log \frac{1}{\delta}\Big)^3 + \log \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\big), \tag{4.13}$$

for some constants $K, c > 0$ and small enough $\delta > 0$. If the prior mass in $\mathscr{F}$ around the true regression function f_0 does not decrease too quickly with decreasing δ, the bound that dominates (4.13) is proportional to the first term on the *r.h.s.*, which leads to near-parametric candidate rates of convergence.

Lemma 4.3. *If there exists a constant $c' > 0$ such that:*

$$\log \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \epsilon\big) \geq -c'\Big(\log \frac{1}{\epsilon}\Big)^3, \tag{4.14}$$

for small enough $\epsilon > 0$, then the prior-mass condition (4.4) is satisfied by the sequence (4.11) for large enough n.

Proof Condition (4.14) implies that (4.13) holds with the lower bound on the *r.h.s.* replaced by only its first term with a larger choice for the constant c. The substitution $\epsilon = K\delta \log(1/\delta)$ leads to a constant and a $\log\log(1/\delta)$ correction, both of which are dominated by $\log(1/\delta)$ for small enough δ. (See the proof of lemma 4.4, where a similar step is made.) It follows that:

$$\log \Pi\big(B(\epsilon; P_0)\big) \geq -c''\Big(\log \frac{1}{\epsilon}\Big)^3,$$

for some constant $c'' > 0$ and small enough ϵ. The remainder of the proof is identical to that of lemma 4.1. □

However, it is also possible that the prior mass around f_0 in the regression class decreases more quickly than (4.14). In that case the lower bound on the *r.h.s.* of (4.13) is determined by the prior on $\mathscr{F}$. The following lemma assumes a so-called *net-prior* on the regression class $\mathscr{F}$, a construction that is explained in subsection 4.5.2.

Lemma 4.4. *For an errors-in-variables model $\mathscr{P}$ based on a regression class $C_{\beta,M}[-A, A]$ with a net-prior Π, the prior-mass condition (4.4) is satisfied by the sequence:*

$$\epsilon_n = n^{-\frac{\beta}{2\beta+1}} (\log n)^{\frac{1}{2\beta}}, \tag{4.15}$$

for large enough n.

Proof Given β, the prior mass in neighbourhoods of the true regression function f_0 for a net prior Π is lower bounded by the expression on the *r.h.s.* in (4.38). Since this term dominates in the *r.h.s.* of (4.13) for small δ, the prior mass of Kullback-Leibler neighbourhoods of P_0 in $\mathscr{P}$ satisfies the following lower bound:

$$\log \Pi\Big(B\big(K\delta \log(1/\delta); P_0\big)\Big) \geq -L \frac{1}{\delta^{1/\beta}},$$

for some constants $K, L > 0$ and small enough δ. Define $\epsilon = K\delta \log(1/\delta)$ and note that, for small enough δ:

$$\begin{aligned}
\frac{1}{\epsilon^{1/\beta}} \Big(\log \frac{1}{\epsilon}\Big)^{1/\beta} &= K^{-1/\beta} \frac{1}{\delta^{1/\beta}} \Big(\log \frac{1}{\delta}\Big)^{-1/\beta} \Big(\log \frac{1}{\delta} - \log K - \log\log \frac{1}{\delta}\Big)^{1/\beta} \\
&\geq K^{-1/\beta} \frac{1}{\delta^{1/\beta}} \Big(\log \frac{1}{\delta}\Big)^{-1/\beta} \Big(\tfrac{1}{2} \log \frac{1}{\delta}\Big)^{1/\beta} \\
&\geq \big(\tfrac{1}{2}\big)^{1/\beta} K^{-1/\beta} \frac{1}{\delta^{1/\beta}}.
\end{aligned}$$

For the first inequality in the above display, we have used that $\log K \leq \log\log \frac{1}{\delta} \leq \frac{1}{4} \log \frac{1}{\delta}$ (for small enough δ). We see that there exists a constant $L' > 0$, such that, for small enough $\epsilon > 0$:

$$\log \Pi\big(B(\epsilon; P_0)\big) \geq -L' \frac{1}{\epsilon^{1/\beta}} \Big(\log \frac{1}{\epsilon}\Big)^{1/\beta}.$$

The sequence ϵ_n satisfies $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$. Define the sequence $a_n = n^{-\beta/(2\beta+1)}$ and note that $\epsilon_n \geq a_n$ (for large enough n) so that for some constant $R > 0$:

$$\begin{aligned}
\log \Pi\big(B(\epsilon_n; P_0)\big) \geq \log \Pi\big(B(a_n; P_0)\big) &\geq -L' \frac{1}{a_n^{1/\beta}} \Big(\log \frac{1}{a_n}\Big)^{1/\beta} \\
&= -Rn^{\frac{1}{2\beta+1}} (\log n)^{\frac{1}{\beta}} = -Rn\epsilon_n^2,
\end{aligned}$$

for large enough n. □

In the case of a parametric regression class ($\mathscr{F}_\Theta$ as defined under case *(iii)* in the beginning of this section) and a prior on Θ with strictly positive and continuous density, the conditions of lemmas 4.1 and 4.3 are satisfied. From lemma 4.15, we know that in the case of a parametric class of regression functions, covering numbers satisfy (4.10). Furthermore, from lemma 4.17, we know that for a parametric class, the prior mass in neighbourhoods of f_0 satisfies (4.14). The resulting conclusion for the rate of convergence in parametric regression classes is given in the theorem below.

We summarize the main results in the following theorem by stating the rates of convergence for the classes defined in the beginning of this section. The proof consists of combination of the preceding lemmas.

Theorem 4.2. *For the specified regression classes, the assertion of theorem 4.1 holds with the following rates of convergence.*

(i) If $\mathscr{F} = \mathrm{Lip}_M(\alpha)$ (for some $\alpha \in (0,1]$ and $M > 0$) with a net prior, the prior-mass condition for neighbourhoods of f_0 in the regression class determines the rate, given by the sequence ϵ_n defined in lemma 4.4 with $\beta = \alpha$:

$$\epsilon_n = n^{-\frac{\alpha}{2\alpha+1}} (\log n)^{\frac{1}{2\alpha}}.$$

(ii) If $\mathscr{F} = D_{\alpha,M}(q)$ (for some $M > 0$ and integer $q \geq 1$) with a net prior, the prior-mass condition for neighbourhoods of f_0 again determines the rate, given by the sequence ϵ_n defined in lemma 4.4 with $\beta = q + \alpha$:

$$\epsilon_n = n^{-\frac{q+\alpha}{2q+2\alpha+1}} (\log n)^{\frac{1}{2q+2\alpha}}.$$

(iii) If $\mathscr{F} = \mathscr{F}_\Theta$ is a parametric class with a prior that has a continuous and strictly positive density throughout Θ, the rate is determined by the posterior convergence with regard to the parameter F and is given by:

$$\epsilon_n = n^{-1/2} (\log n)^{3/2}.$$

Concerning the parametric rate of convergence, it is stressed that this rate applies to the full, non-parametric problem and can not be compared with semi-parametric rates for estimation of the parameter θ in the presence of the nuisance parameter F. With regard to the logarithmic corrections to the powers of n in the expressions for the rate of convergence in Lipschitz- and smoothness-classes, we note that they originate from (the proof of) lemma 4.4: the logarithm is introduced by the transition from δ to ϵ, which compensates for the logarithmic correction in the extent of the Kullback-Leibler neighbourhoods $B(K\delta \log(1/\delta); P_0)$. When considering near-parametric rates (as in lemmas 4.1 and 4.3), logarithmic corrections of this kind do not influence the calculation, but they do play a role in non-parametric regression. It is possible that these logarithmic corrections to the rate can be omitted, the proof

depending on a version of theorem 4.1 along the lines of theorem 2.4 of Ghosal *et al.* (2000) [39], in which the prior-mass condition is replaced by a more complicated, but less demanding bound on a ratio of prior masses. Note that the rate (4.15) approaches that given in (4.12) for large values of β, *i.e.* for regression classes with a high degree of differentiability.

Regarding classes with a high degree of differentiabilty, one might expect that suitably restricted classes of analytic regression functions would allow for convergence at the rate (4.15) in the limit $\beta \to \infty$, *i.e.* $1/\sqrt{n}$. However, in that case (4.9) and (4.13) are dominated by the contribution from the parameter $F \in D$, so the expected result would be the parametric rate of convergence given above, *i.e.* $1/\sqrt{n}$ with logarithmic correction of the order $(\log n)^{3/2}$.

4.3 Model entropy

One of the two primary conditions in theorems on non-parametric Bayesian rates of convergence (see, *e.g.* theorem 4.1), is an upper-bound on the covering numbers with respect to a metric on the model, in our case the Hellinger metric. In this section, we relate the Hellinger metric entropy of the model to entropy numbers of the three parametrizing spaces, *i.e.* I, $\mathscr{F}$ and D. Due to technical reasons (see subsection 4.3.3, which contains the proofs of all lemmas in this section), we can and shall express most results in terms of the $L_1(\mu)$-norm rather than the Hellinger metric, demonstrating in the (proof of) theorem 4.3 that this does not influence the entropy calculation.

4.3.1 Nets in parametrizing spaces

We start the discussion by considering the $L_1(\mu)$-distance between densities in the model that differ only in one of the three parameters (σ, f, F), the goal being the definition of an ϵ-net over $\mathscr{P}$ from ϵ-nets over the spaces I, $\mathscr{F}$ and D separately.

With the following lemma, we indicate the possibility of generalizing the discussion that follows to situations in which less is known about the error distribution, by a bound on the $L_1(\mu)$-difference under variation of the parameter for the error distribution. For the next lemma only, we define $\{\psi_\sigma : \sigma \in \Sigma\}$ to be a family of Lebesgue densities of probability distributions on $\mathbb{R}^2$, parametrized by σ in some (parametric or non-parametric) set Σ. The densities $p_{\sigma,f,F}$ are still given by a convolution *c.f.* (4.2) (because we maintain the assumption of independence of Z and (e, f)).

Lemma 4.5. *For every $f \in \mathscr{F}$ and $F \in D$,*

$$\|p_{\sigma,f,F} - p_{\tau,f,F}\|_{1,\mu} \leq \|\psi_\sigma - \psi_\tau\|_{1,\mu},$$

for all $\sigma, \tau \in \Sigma$.

Specializing back to the situation of interest, we find the following lemma.

Lemma 4.6. *In the case of equally distributed, independent normal errors* (e_1, e_2) *with mean zero and equal but unknown variance in the interval* $[\underline{\sigma}, \overline{\sigma}]$*:*

$$\|\psi_\sigma - \psi_\tau\|_{1,\mu} \leq 4\overline{\sigma}\underline{\sigma}^{-2}|\sigma - \tau|.$$

Similar inequalities can be derived for other parametric families of kernels, for instance the Laplace kernel. In the case of a non-parametric family of error distributions, it may be necessary to derive a (sharper) bound, based on the Hellinger distance between $p_{\sigma,f,F}$ and $p_{\tau,f,F}$. This generalized approach is not pursued here and the rest of this paper relies on the assumption that the errors (e_1, e_2) are as in the above lemma.

Next we consider the dependence of densities in the model on the regression function f.

Lemma 4.7. *There exists a constant* $K > 0$ *such that for all* $\sigma \in I$ *and all* $F \in D[-A, A]$*:*

$$\|p_{\sigma,f,F} - p_{\sigma,g,F}\|_{1,\mu} \leq K\|f - g\|_{1,F}, \tag{4.16}$$

for all $f, g \in \mathscr{F}$*.*

The bound depends on the distribution F for the underlying random variable Z and proves the claim we made earlier, concerning identifiability of the regression function only up to null-sets of the distribution F. To derive a bound that is independent of F, we note that for all $F \in D$ and all $f, g \in C[-A, A]$:

$$\|f - g\|_{1,F} \leq \sup\big\{|f - g|(z) \,:\, z \in [-A, A]\big\} = \|f - g\|, \tag{4.17}$$

the right side being finite as a result of continuity of f and g and compactness of the interval $[-A, A]$. Note that we cannot simply equate the uniform norm $\|\,.\,\|$ in (4.17) to the L_∞-norm because the Lebesgue measure on $[-A, A]$ does not dominate all $F \in D$.

The bound $H^2(P, Q) \leq \|p - q\|_{1,\mu}$ suggests that metric entropy numbers for the Hellinger distance can safely be upper-bounded by those for the $L_1(\mu)$-norm. In cases where the class of regression functions is non-parametric and in fact large enough to dominate the metric entropy of the model, this line of reasoning is insufficient for optimal rates of convergence in the Hellinger distance. The reason is the fact that it is the *squared* Hellinger distance that is dominated by the $L_1(\mu)$-distance and not the Hellinger distance itself. As long as $L_1(\mu)$ entropy numbers are logarithmic, transition from $L_1(\mu)$- to Hellinger coverings leads only to a larger constant. However, if the small-ϵ behaviour of $L_1(\mu)$ entropy numbers is dominated by terms of the form (4.29)), the replacement of ϵ by ϵ^2 influences the calculation. Therefore, we also provide the following lemma.

Lemma 4.8. *For all* $\sigma \in I$*,* $f, g \in \mathscr{F}$ *and* $F \in D$*:*

$$H(P_{\sigma,f,F}, P_{\sigma,g,F}) \leq \frac{1}{2\sigma}\Big(\int_{[-A,A]} \big(f(z) - g(z)\big)^2\, dF(z)\Big)^{1/2}.$$

Although useful, the above bound depends on the particular values of σ, F, which is undesirable in situations below. The lower bound for the interval I and the uniform bound on $|f-g|(z)$ serve to prove a bound on the Hellinger distance proportional to the uniform norm (as opposed to its square-root) of the difference between regression parameters.

Corollary 4.1. *There exists a constant $L > 0$ such that for all $\sigma \in I$, $f, g \in \mathscr{F}$ and $F \in D$:*

$$H(P_{\sigma,f,F}, P_{\sigma,g,F}) \leq L\|f-g\|. \tag{4.18}$$

The above two lemmas and the fact that approximation in the uniform norm of subclasses of bounded continuous functions on closed intervals is well-understood, strongly suggests that the class of regression functions is to be endowed with the uniform norm to find nets. We do this in subsection 4.5.1 for the regression classes mentioned earlier.

To bound the contribution of the parameter F to the covering numbers of the model, we approximate F by a discrete distribution F' with a number of support points that is bounded by the approximation error in $L_1(\mu)$. Note that the number of support points needed depends on a power of $\log(1/\epsilon)$, so that a sharper bound in terms of the Hellinger distance is not necessary (see above).

Lemma 4.9. *There exist constants $C, C' > 0$ such that for all $(\sigma, f) \in I \times \mathscr{F}$ and $F \in D$, there is a discrete F' on $[-A, A]$ with less than $C(\log(1/\epsilon))^2$ support points such that*

$$\|p_{\sigma,f,F} - p_{\sigma,f,F'}\|_{1,\mu} \leq C'\epsilon.$$

We stress that the particular choice F' depends on the regression function f. The above lemma implies that the set D_ϵ of all discrete $F \in D$ with less than $C(\log(1/\epsilon))^2$ support points parametrizes an ϵ-net over $\mathscr{P}$. For any fixed pair $(\sigma, f) \in I \times \mathscr{F}$, the ϵ-net parametrized by D_ϵ is a 2ϵ-net over the submodel $\mathscr{P}_{\sigma,f} = \{p_{\sigma,f,F} \in \mathscr{P} : F \in D\}$ so that

$$N\big(\epsilon, \mathscr{P}_{\sigma,f}, \|\,.\,\|_{1,\mu}\big) \leq N\big(2\epsilon, \{p_{\sigma,f,F} \in \mathscr{P} : F \in D_\epsilon\}, \|\,.\,\|_{1,\mu}\big).$$

The direct nature of the above approximation (as opposed to the procedure for the parameters σ and f, where we first bound by a norm on the parametrizing variable and then calculate the entropy in the parametrizing space) circumvents the notoriously difficult dependence of mixture densities on their mixing distribution, responsible for the (logarithmically) slow rate of convergence in deconvolution problems. Indeed, problems of this nature plague the method of Fan and Truong (1993) [31], which is based on a kernel-estimate for F and leads to a Nadaraya-Watson-type of estimator for the regression function. Here we are only interested in covering the model $\mathscr{P}$, which allows us to by-pass the deconvolution problem by means of the above lemma.

4.3.2 Metric entropy of the errors-in-variables model

This subsection is devoted entirely to the following theorem, which uses the lemmas of the previous subsection to calculate the Hellinger entropy of the errors-in-variables model $\mathscr{P}$.

Theorem 4.3. *Suppose that the regression family $\mathscr{F}$ is one of those specified in the beginning of section 4.2). Then there exist constants $L, L' > 0$ such that the Hellinger covering numbers of the model $\mathscr{P}$ satisfy:*

$$\log N(\epsilon, \mathscr{P}, H) \leq L'\Big(\log \frac{1}{\epsilon}\Big)^3 + \log N(L\epsilon, \mathscr{F}, \|\,.\,\|), \tag{4.19}$$

for small enough ϵ.

Proof If the class of regression functions $\mathscr{F}$ is a Lipschitz-class with exponent in $(0,1)$, we set α equal to that exponent. In other cases we set $\alpha = 1$.

Let $\epsilon > 0$ be given, fix some $\sigma \in I$, $f \in \mathscr{F}$. According to lemma (4.9) the collection $\mathscr{P}^\epsilon_{\sigma,f}$ of all $p_{\sigma,f,F'}$ where F' is a discrete distribution in D with at most $N_\epsilon = \alpha^2 C\big(\log(1/\epsilon)\big)^2$ support points, forms an ϵ^α-net over $\mathscr{P}_{\sigma,f}$ with respect to the $L_1(\mu)$-norm. Therefore any ϵ^α-net $\mathscr{Q}^\epsilon_{\sigma,f}$ over $\mathscr{P}^\epsilon_{\sigma,f}$ is a $2\epsilon^\alpha$-net over $\mathscr{P}_{\sigma,f}$. Let $\mathscr{S}_\epsilon$ be a minimal ϵ^α-net for the simplex with ℓ_1-norm in $\mathbb{R}^{N_\epsilon}$. As is shown by lemma A.4 in Ghosal and Van der Vaart (2001) [40], the order of $\mathscr{S}_\epsilon$ does not exceed $(5/\epsilon^\alpha)^{N_\epsilon}$. Next we define the grid $G_\epsilon = \{0, \pm\epsilon, \pm 2\epsilon, \ldots\} \subset [-A, A]$ and $\mathscr{Q}^\epsilon_{\sigma,f}$ as the collection of all distributions on $[-A, A]$ obtained by distributing the weights in a vector from $\mathscr{S}_\epsilon$ over the points in G_ϵ. We project an arbitrary $p_{\sigma,f,F'}$ in $\mathscr{P}^\epsilon_{\sigma,f}$ onto $\mathscr{Q}^\epsilon_{\sigma,f}$ in two steps: given that $F' = \sum_{i=1}^{N_\epsilon} \lambda_i \delta_{z_i}$, for some set of N_ϵ points $z_i \in [-A, A]$ and non-negative weights such that $\sum_i \lambda_i = 1$, we first project the vector λ onto a vector in $\mathscr{S}_\epsilon$ and second, shift the resulting masses to the closest point in G_ϵ. One easily sees that the first step leads to a new distribution F'' such that:

$$\|p_{\sigma,f,F'} - p_{\sigma,f,F''}\|_{1,\mu} \leq \epsilon.$$

As for the second step, in which $F'' = \sum_{i=1}^{N_\epsilon} \lambda'_i \delta_{z_i}$ is 'shifted' to a new distribution $F''' = \sum_{i=1}^{N_\epsilon} \lambda'_i \delta_{z'_i}$ such that $|z_i - z'_i| \leq \epsilon$, we note that:

$$\begin{aligned}
\big|p_{\sigma,f,F''} - p_{\sigma,f,F'''}\big|(x,y) &\leq \sum_{i=1}^{N_\epsilon} \lambda'_i \big|\varphi_\sigma(x - z_i)\varphi_\sigma(y - f(z_i)) - \varphi_\sigma(x - z'_i)\varphi_\sigma(y - f(z'_i))\big| \\
&\leq \sum_{i=1}^{N_\epsilon} \lambda'_i \Big(\big|\varphi_\sigma(x - z_i) - \varphi_\sigma(x - z'_i)\big|\varphi_\sigma(y - f(z_i)) \\
&\qquad + \big|\varphi_\sigma(y - f(z_i)) - \varphi_\sigma(y - f(z'_i))\big|\varphi_\sigma(x - z'_i)\Big),
\end{aligned}$$

which implies that the $L_1(\mu)$-difference satisfies:

$$\begin{aligned}
&\|p_{\sigma,f,F''} - p_{\sigma,f,F'''}\|_{1,\mu} \\
&\qquad \leq \sum_{i=1}^{N_\epsilon} \lambda'_i \Big(\int \big|\varphi_\sigma(x - z_i) - \varphi_\sigma(x - z'_i)\big|\, dx + \int \big|\varphi_\sigma(y - f(z_i)) - \varphi_\sigma(y - f(z'_i))\big|\, dy \Big).
\end{aligned}$$

By assumption, the family of regression functions satisfies (4.7), which is used to establish that there exists a constant $K > 0$ such that

$$\|p_{\sigma,f,F''} - p_{\sigma,f,F'''}\|_{1,\mu}, \leq K\epsilon^{\alpha},$$

(for small enough ϵ), along the same lines as the proof of lemma 4.7. Summarizing, we assert that for some constant $K_3 > 0$, $\mathscr{Q}^{\epsilon}_{\sigma,f}$ is a $K_3^2\epsilon^{\alpha}$-net over $\mathscr{P}_{\sigma,f}$. There exist an ϵ^{α}-net I_{ϵ} over I (with norm equal to absolute differences) and an $\epsilon^{\alpha/2}$-net $\mathscr{F}_{\epsilon}$ over $\mathscr{F}$ in the uniform norm. (The order of $\mathscr{F}_{\epsilon}$ is bounded in lemmas 4.13 and 4.15.) By virtue of the triangle inequality and with the help of lemma 4.5 and corollary 4.1, we find that constants $K_1, K_2 > 0$ exist such that:

$$\begin{aligned} H(P_{\sigma,f,F}, P_{\tau,g,F'}) &\leq H(P_{\sigma,f,F}, P_{\tau,f,F}) + H(P_{\tau,f,F}, P_{\tau,g,F}) + H(P_{\tau,g,F}, P_{\tau,g,F'}) \\ &\leq \|p_{\sigma,f,F} - p_{\tau,f,F})\|^{1/2}_{1,\mu} + K\|f - g\| + \|p_{\tau,g,F} - p_{\tau,g,F'}\|^{1/2}_{1,\mu} \\ &\leq K_1|\sigma - \tau|^{1/2} + K_2\|f - g\| + \|p_{\tau,g,F} - p_{\tau,g,F'}\|^{1/2}_{1,\mu}, \end{aligned}$$

for all $\sigma \in I$, $\tau \in I_{\epsilon}$, $f \in \mathscr{F}$, $g \in \mathscr{F}_{\epsilon}$ and $F, F' \in D$. For every fixed pair $(\tau, g) \in I_{\epsilon} \times \mathscr{F}_{\epsilon}$, we define the $K_3^2\epsilon^{\alpha}$-net $\mathscr{Q}^{\epsilon}_{\tau,g}$ like above and choose F' in the above display so that $p_{\tau,g,F'}$ lies in $\mathscr{Q}^{\epsilon}_{\tau,g}$ and approximates $p_{\tau,g,F}$ to within $L_1(\mu)$-distance proportional to ϵ^{α}. This shows that the set:

$$\mathscr{Q}_{\epsilon} = \bigcup\{\mathscr{Q}^{\epsilon}_{\tau,g} : \tau \in I_{\epsilon}, g \in \mathscr{F}_{\epsilon}\},$$

forms a $K\epsilon^{\alpha/2}$-net over $\mathscr{P}$ with respect to the Hellinger distance, where $K = K_1 + K_2 + K_3$. The order of this net can be calculated and forms an upper bound for the Hellinger covering number of the model.

$$\log N(K\epsilon^{\alpha/2}, \mathscr{P}, H) \leq \log N(\epsilon^{\alpha}, I, |\,.\,|) + \log N(\epsilon^{\alpha/2}, \mathscr{F}, \|\,.\,\|) + \log N(\mathscr{Q}^{\epsilon}_{\tau,g}),$$

where $N(\mathscr{Q}^{\epsilon}_{\tau,g})$ denotes the uniform bound on the number of points in the nets $\mathscr{Q}^{\epsilon}_{\tau,g}$, given by:

$$\log N(\mathscr{Q}^{\epsilon}_{\tau,g}) = L''\Big(\log\frac{1}{\epsilon}\Big)^3,$$

for some constant $L'' > 0$ as is easily checked from the above. Moreover, the covering numbers for the finite-dimensional, bounded space I satisfy, for some constant $L''' > 0$:

$$\log N(\epsilon^{\alpha}, I, |\,.\,|) \leq L'''\log\frac{1}{\epsilon}.$$

(Note that in the two displays above, any exponent for ϵ (*e.g.* $\alpha/2$) is absorbed in the constants L' and L''). Note that for small enough ϵ, the contribution from the mixing parameter F dominates that of the parameter σ. Eventually, we find the bound:

$$\log N(\epsilon, \mathscr{P}, H) \leq L'\Big(\log\frac{1}{\epsilon}\Big)^3 + \log N(L\epsilon, \mathscr{F}, \|\,.\,\|),$$

for small enough $\epsilon > 0$ and some $L, L' > 0$. □

4.3.3 Proofs of several lemmas

Proof of lemma 4.5 Fix $f \in \mathscr{F}$ and $F \in D$, let $\sigma, \tau \in \Sigma$ be given. Consider the $L_1(\mu)$ difference:

$$\|p_{\sigma,f,F} - p_{\tau,f,F}\|_{1,\mu} \leq \int_{\mathbb{R}} \int_{\mathbb{R}^2} \Big| \psi_\sigma\big(x - z, y - f(z)\big) - \psi_\tau\big(x - z, y - f(z)\big) \Big| \, d\mu(x,y) \, dF(z),$$

by Fubini's theorem. Translation invariance of the Lebesgue measure and the domain of integration $\mathbb{R}^2$ make it possible to translate over $(z, f(z))$ to render the inner integral independent of z and integrate with respect to F with the following result:

$$\|p_{\sigma,f,F} - p_{\tau,f,F}\|_{1,\mu} \leq \int_{\mathbb{R}^2} \Big| \psi_\sigma(x,y) - \psi_\tau(x,y) \Big| \, d\mu(x,y),$$

thus leading to an upper bound that is independent of both f and F. □

Proof of lemma 4.6 The $L_1(\mu)$-difference of the densities ψ_σ and ψ_τ equals the total-variational difference between the corresponding distributions Ψ_σ and Ψ_τ and can be expressed in terms of the event $\{\psi_\sigma > \psi_\tau\}$ as follows:

$$\|\psi_\sigma - \psi_\tau\|_{1,\mu} = 2 \, \big(\, \Psi_\sigma(\psi_\sigma > \psi_\tau) - \Psi_\tau(\psi_\sigma > \psi_\tau) \, \big).$$

In the case of normally and equally distributed, independent errors (e_1, e_2) the kernel is $\psi_\sigma(x,y) = \varphi_\sigma(x)\varphi_\sigma(y)$, with $\sigma \in I$. Assuming that $\sigma < \tau$, the event in question is a ball in $\mathbb{R}^2$ of radius r_0 centred at the origin (and its complement if $\sigma > \tau$), where $r_0^2 = (2\sigma^2\tau^2/(\tau^2 - \sigma^2)) \log(\tau^2/\sigma^2)$. Integrating the normal kernels over this ball, we find:

$$\|\psi_\sigma - \psi_\tau\|_{1,\mu} = 2 \Big| e^{-\frac{1}{2}(r_0/\sigma)^2} - e^{-\frac{1}{2}(r_0/\tau)^2} \Big| = 2e^{-\frac{1}{2}(r_0/\sigma)^2} \Big| 1 - \frac{\sigma^2}{\tau^2} \Big| \leq \frac{4\overline{\sigma}}{\underline{\sigma}^2} |\sigma - \tau|,$$

where we have used the upper and lower bounds for the interval I. □

Proof of lemma 4.7 Let $\sigma \in I$, $F \in D[-A, A]$ and $f, g \in \mathscr{F}$ be given. Since the x-dependence of the densities $p_{\sigma,f,F}$ and $p_{\sigma,g,F}$ is identical and can be integrated out, the $L_1(\mu)$-difference can be upper-bounded as follows:

$$\|p_{\sigma,f,F} - p_{\sigma,g,F}\|_{1,\mu} \leq \int_{\mathbb{R}} \int_{\mathbb{R}} \big| \varphi_\sigma\big(y - f(z)\big) - \varphi_\sigma\big(y - g(z)\big) \big| \, dy \, dF(z).$$

Fix a $y \in \mathbb{R}$ and $z \in [-A, A]$. We note:

$$\big| \varphi_\sigma(y - f(z)) - \varphi_\sigma(y - g(z)) \big| \leq \Big| \int_{y - f(z)}^{y - g(z)} \varphi'_\sigma(u) \, du \Big| \leq \sup\big\{ \big| \varphi'_\sigma(u) \big| \, : \, u \in J \big\} \, \big| f(z) - g(z) \big|,$$

where $J = [y - f(z) \vee g(z), y - f(z) \wedge g(z)]$. The uniform bound on the functions in the regression class $\mathscr{F}$ guarantees that $J \subset J' = [y - B, y + B]$. If $y \geq 2B$, then $y - B \geq \frac{1}{2}y \geq B > 0$, so if, in addition, $\frac{1}{2}y \geq \overline{\sigma}$, we see that for all $u \in J'$, $u \geq \frac{1}{2}y \geq \sigma$, thus restricting u to the region in which the derivative of the normal density decreases monotonously:

$$\big| \varphi'_\sigma(u) \big| \leq \big| \varphi'_\sigma(\tfrac{1}{2}y) \big|.$$

Symmetry of the normal density allows us to draw the same conclusion if y lies below $-2B$ and $-2\overline{\sigma}$. Using the explicit form of the normal density and the constant $T = 2(B \vee \overline{\sigma})$, we derive the following upper bound on the supremum:

$$\sup\bigl\{\bigl|\varphi'_\sigma(u)\bigr| \,:\, u \in J\bigr\} \leq Ks(y),$$

where the function s is given by:

$$s(y) = \begin{cases} |y|\,\varphi_{2\overline{\sigma}}(y), & \text{if } |y| \geq T, \\ \|\varphi'_{\underline{\sigma}}\|_\infty, & \text{if } |y| < T. \end{cases}$$

Note that s does not depend on the values of the parameters. Therefore:

$$\|p_{\sigma,f,F} - p_{\sigma,g,F}\|_{1,\mu} \leq \int_{\mathbb{R}}\int_{\mathbb{R}} Ks(y)\bigl|f(z) - g(z)\bigr|\,dy\,dF(z).$$

Since the integral over $s(y)$ is finite, the asserted bound follows. □

Proof of lemma 4.8 Consider a binary experiment $E_1 = (\mathbb{R}^3, \mathscr{B}^{(3)}, \{P, Q\})$, giving two possible distributions P, Q for the triplet (X, Y, Z) that describes the errors-in-variables model (*c.f.* (4.1)). The map T that projects by $T(X, Y, Z) = (X, Y)$ leads to another binary experiment $E_2 = (\mathbb{R}^2, \mathscr{B}^{(2)}, \{P^T, Q^T\})$ which is less informative[1] than E_1. This property follows from the fact that $\sigma(X, Y) \subset \mathscr{B}^{(2)}$ is such that $T^{-1}\bigl(\sigma(X, Y)\bigr) \subset \sigma(X, Y, Z) \subset \mathscr{B}^{(3)}$, which makes it possible to identify every test function in E_2 with a test function in E_1, while there may exist test functions on $\mathbb{R}^3$ that are not measurable with respect to $T^{-1}\bigl(\sigma(X, Y)\bigr)$. Corollary 17.3 in Strasser (1985) [85] asserts that the Hellinger distance decreases when we make the transition from a binary experiment to a less informative binary experiment, so we see that:

$$H(P^T, Q^T) \leq H(P, Q). \tag{4.20}$$

In the case at hand, we choose $P^T = P_{\sigma,f,F}$ and $Q^T = P_{\sigma,g,F}$. From the definition of the errors-in-variables model (4.1), we obtain the conditional laws:

$$\begin{aligned} \mathscr{L}_P\bigl(\, X, Y \bigm| Z \,\bigr) &= N(Z, \sigma^2) \times N(f(Z), \sigma^2), \\ \mathscr{L}_Q\bigl(\, X, Y \bigm| Z \,\bigr) &= N(Z, \sigma^2) \times N(g(Z), \sigma^2), \end{aligned}$$

and, of course, $\mathscr{L}_P(Z) = \mathscr{L}_Q(Z) = F$. It follows that:

$$\begin{aligned} H^2(P, Q) &= \int_{\mathbb{R}^3} \bigl(dP^{1/2} - dQ^{1/2}\bigr)^2 \\ &= \int_{\mathbb{R}^3} \varphi_\sigma(x - z)\Bigl(\varphi_\sigma\bigl(y - f(z)\bigr)^{1/2} - \varphi_\sigma\bigl(y - g(z)\bigr)^{1/2}\Bigr)^2 dF(z)\,dx\,dy \\ &= \int_{[-A,A]} H^2\bigl(N(f(z), \sigma^2), (N(g(z), \sigma^2)\bigr)\,dF(z), \end{aligned}$$

[1] The phrase "less informative" is defined in the sense of Le Cam, *i.e.* for every test function ϕ_2 in E_2, there exists a test function ϕ_1 in E_1 such that $P\phi_1 \leq P^T\phi_2$ and $Q\phi_1 \geq Q^T\phi_2$ (see, for instance, Strasser (1985) [85], definition 15.1).

by Fubini's theorem. A straightforward calculation shows that:

$$H^2\big(N(f(z),\sigma^2),(N(g(z),\sigma^2)\big) = 2\Big(1 - e^{-\frac{1}{2}\left(f(z)-g(z)\right)^2/(2\sigma)^2}\Big) \leq \frac{1}{4\sigma^2}\big(f(z)-g(z)\big)^2,$$

where we use that $1 - e^{-x} \leq x$ for all $x \geq 0$. Upon combination of the above two displays and (4.20), we obtain:

$$H^2(P_{\sigma,f,F}, P_{\sigma,g,F}) \leq \frac{1}{4\sigma^2}\int_{[-A,A]}\big(f(z)-g(z)\big)^2\,dF(z),$$

which proves the assertion. □

Proof of lemma 4.9 Let $\epsilon > 0$, $\sigma \in I$, $f \in \mathscr{F}$ be given, fix $M \geq 2A \vee 2B$ and $k \geq 1$. A Taylor-expansion up to order $k-1$ of the exponential in the normal density demonstrates that:

$$\begin{aligned}\Big|\varphi_\sigma(x-z) - \frac{1}{\sigma\sqrt{2\pi}}\sum_{j=0}^{k-1}\frac{1}{j!}(-\tfrac{1}{2})^j\Big(\frac{x-z}{\sigma}\Big)^{2j}\Big| &\leq \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{k!}(\tfrac{1}{2})^k\Big(\frac{x-z}{\sigma}\Big)^{2k}\\ &\leq \frac{1}{\sigma\sqrt{2\pi}}\Big(\frac{e}{2k}\Big)^k\Big(\frac{x-z}{\sigma}\Big)^{2k},\end{aligned}$$

where we have used that $k! \geq k^k e^{-k}$. Similarly, we obtain:

$$\Big|\varphi_\sigma\big(y-f(z)\big) - \frac{1}{\sigma\sqrt{2\pi}}\sum_{j=0}^{k-1}\frac{1}{j!}(-\tfrac{1}{2})^j\Big(\frac{y-f(z)}{\sigma}\Big)^{2j}\Big| \leq \frac{1}{\sigma\sqrt{2\pi}}\Big(\frac{e}{2k}\Big)^k\Big(\frac{y-f(z)}{\sigma}\Big)^{2k}.$$

Considering $|x|, |y| \leq M$ and using that $\sigma \geq \overline{\sigma} > 0$, we see that there exists a constant $C_1 > 0$ (independent of σ and f) such that both residuals of the last two displays are bounded above by $(C_1M^2/k)^k$. So for all x, y like above,

$$\begin{aligned}&\big|p_{\sigma,f,F} - p_{\sigma,f,F'}\big|(x,y)\\ &\quad\leq \frac{1}{2\pi\sigma^2}\Big|\int\sum_{i,j=0}^{k-1}\frac{1}{i!j!}(-\tfrac{1}{2})^{i+j}\Big(\frac{x-z}{\sigma}\Big)^{2i}\Big(\frac{y-f(z)}{\sigma}\Big)^{2j}d(F-F')(z)\Big| \qquad (4.21)\\ &\qquad + 4\Big(\frac{C_1M^2}{k}\Big)^k + \Big(\frac{C_1M^2}{k}\Big)^{2k}.\end{aligned}$$

Lemma A.1 in Ghosal and Van der Vaart (2001) [40] asserts that there exists a discrete distribution F' on $[-A, A]$ with at most (k^2+1) support points such that for all functions $\psi_{f,ij}(z) = z^{2i}f^{2j}(z)$ the F- and F'-expectations coincide, *i.e.*:

$$\int_{[-A,A]}\psi_{f,ij}\,dF = \int_{[-A,A]}\psi_{f,ij}\,dF'.$$

Thus choosing F', the first term in (4.21) vanishes and we see that (for large enough k):

$$\sup_{|x|\vee|y|\leq M}\big|p_{\sigma,f,F} - p_{\sigma,f,F'}\big|(x,y) \leq 5\Big(\frac{C_1M^2}{k}\Big)^k. \qquad (4.22)$$

For points (x, y) outside $[-M, M] \times [-M, M]$, we note that there exists a constant $C_2 > 0$ such that for all $|x| \geq 2A$, $|y| \geq 2B$:

$$\varphi_\sigma(x - z) \leq \varphi_\sigma\big(\tfrac{x}{2}\big) \leq C_2 \varphi_{\overline{\sigma}}\big(\tfrac{x}{2}\big),$$
$$\varphi_\sigma\big(y - f(z)\big) \leq \varphi_\sigma\big(\tfrac{y}{2}\big) \leq C_2 \varphi_{\overline{\sigma}}\big(\tfrac{y}{2}\big),$$

($C_2 = \|\varphi_{\underline{\sigma}}\|_\infty / \|\varphi_{\overline{\sigma}}\|_\infty$ will do). Since $M \geq 2A \vee 2B$, there exists a constants $C_3, C_4 > 0$ such that:

$$\begin{aligned}\int_{|x|\vee|y|>M} p_{\sigma,f,F}(x,y)\,d\mu(x,y) &\leq C_2 \int_{|x|>M} \varphi_{\overline{\sigma}}\big(\tfrac{x}{2}\big)\,dx \int\int \varphi_\sigma\big(y - f(z)\big)\,dF(z)\,dy \\ &\quad + C_2 \int_{|y|>M} \varphi_{\overline{\sigma}}\big(\tfrac{y}{2}\big)\,dy \int\int \varphi_\sigma(x - z)\,dF(z)\,dx \\ &= 4C_2 \int_{x>M} \varphi_{\overline{\sigma}}\big(\tfrac{x}{2}\big)\,dx \leq 4C_2 \int_{x>M} \frac{x}{M} \varphi_{\overline{\sigma}}\big(\tfrac{x}{2}\big)\,dx \\ &\leq C_3 e^{-C_4 M^2},\end{aligned} \tag{4.23}$$

where we have used Fubini's theorem and translation invariance of Lebesgue measure in the second step and the fact that $\varphi'_\sigma(x) = -(x/\sigma^2)\varphi_\sigma(x)$ in the last. Now, let $\epsilon > 0$ be given. We decompose the domain of integration for the $L_1(\mu)$-difference between $p_{\sigma,f,F}$ and $p_{\sigma,f,F'}$ into the region where $|x| \vee |y| \leq M$ and its complement. Using the uniform bound (4.22) on the region bounded by M and (4.23) for the tails, we find that there is a constant D_1 such that:

$$\|p_{\sigma,f,F} - p_{\sigma,f,F'}\|_{1,\mu} \leq D_1 \Big(M^2 \Big(\frac{C_1 M^2}{k} \Big)^k + e^{-C_4 M^2} \Big). \tag{4.24}$$

In order to bound the *r.h.s.* by ϵ we fix M in terms of ϵ:

$$M = \sqrt{\tfrac{1}{C_4} \log \tfrac{1}{\epsilon}},$$

and note that the lower bound $M \geq 2A \vee 2B$ is satisfied for small enough ϵ. Upon substitution, the first term in (4.24) leads to $(D_1/C_4) D_2^k\, e^{(k+1)\log\log \frac{1}{\epsilon}}\, e^{-k \log k}$ (where $D_2 = C_1/C_4$), so that the choice:

$$k \geq D_3 \log \tfrac{1}{\epsilon},$$

(for some large $D_3 > D_2$) suffices to upper bound the $L_1(\mu)$-difference appropriately. The smallest integer k above the indicated bound serves as the minimal number of support points needed. □

Note that the f-dependence of the functions $\psi_{f,ij}$ carries over to the choice for F', which is therefore f-dependent as well.

4.4 Model prior

Assume that the model is well-specified and denote by $P_0 \in \mathscr{P}$ (corresponding to some, not necessarily unique, $\sigma_0 \in I$, $f_0 \in \mathscr{F}$ and $F_0 \in D$) the true distribution underlying the *i.i.d.*

sample. We define a prior Π on $\mathscr{P}$ by defining priors on the parameter spaces I, $\mathscr{F}$ and D and taking Π equal to the probability measure induced by the map $(\sigma, f, F) \mapsto P_{\sigma,f,F}$ from $I \times \mathscr{F} \times D$ with product-measure to $\mathscr{P}$. The prior on I is denoted Π_I and is assumed to have a density π_I, continuous and strictly positive at σ_0. The prior $\Pi_{\mathscr{F}}$ on $\mathscr{F}$ is specified differently for each of the classes defined in the beginning of section 4.2, but all have as their domain the Borel σ-algebra generated by the norm topology on $C[-A, A]$. The definition of these priors is postponed to subsection 4.5.2. The prior Π_D on D is based on a Dirichlet process with base measure α which has a continuous and strictly positive density on all of $[-A, A]$. The domain of Π_D is the Borel σ-algebra generated by the topology of weak convergence.

The fact that these priors are defined on the product of the parameter spaces rather than the errors-in-variables model $\mathscr{P}$ itself, necessitates a lemma asserting appropriate measurability. So before we discuss the properties of priors, we show that the map $\hat{p}$ that takes parameters (σ, f, F) into densities $p_{\sigma,f,F}$ (*c.f.* (4.2)) is measurable.

Lemma 4.10. *Endow I and $\mathscr{F}$ with their norm topology and D with the topology of weak convergence. Then the map $\hat{p} : I \times \mathscr{F} \times D \to L_1(\mu)$ is continuous in the product topology.*

Proof The space D with the topology of weak convergence is metric, so the product topology on $I \times \mathscr{F} \times D$ is a metric topology as well. Let (σ_n, f_n, F_n) be a sequence, converging to some point (σ, f, F) in $I \times \mathscr{F} \times D$ as $n \to \infty$. As a result of the triangle inequality and lemmas 4.5–4.7, the $L_1(\mu)$-distance satisfies:

$$\big\|p_{\sigma_n,f_n,F_n} - p_{\sigma,f,F}\big\|_{1,\mu} \leq K_1|\sigma_n - \sigma| + K_2\|f_n - f\| + \big\|p_{\sigma,f,F_n} - p_{\sigma,f,F}\big\|_{1,\mu}, \tag{4.25}$$

for some constants $K_1, K_2 > 0$. Since F_n converges to F weakly, the continuity of the regression function f, combined with the continuity and boundedness of the Gaussian kernel and the portmanteau lemma guarantee that

$$\int_{[-A,A]} \varphi_\sigma\big(x - z\big)\varphi_\sigma\big(y - f(z)\big)\, dF_n(z) \to \int_{[-A,A]} \varphi_\sigma\big(x - z\big)\varphi_\sigma\big(y - f(z)\big)\, dF(z),$$

as $n \to \infty$ for all $(x, y) \in \mathbb{R}^2$. Using the (μ-integrable) upper-envelope for the model $\mathscr{P}$ and dominated convergence, we see that

$$\big\|p_{\sigma,f,F_n} - p_{\sigma,f,F}\big\|_{1,\mu} \to 0,$$

and hence the *r.h.s.* of (4.25) goes to zero. We conclude that $\hat{p}$ is continuous in the product topology. □

Note that the $L_1(\mu)$- and Hellinger topologies on the model $\mathscr{P}$ are equivalent, so that the above lemma implies continuity of $\hat{p}$ in the Hellinger topology. Hence $\hat{p}^{-1}$ is a well-defined map between the Borel σ-algebras of the model with the Hellinger topology and the product $I \times \mathscr{F} \times D$.

The following lemma establishes that the prior-mass condition (4.4) can be analysed for the regression class and the parameter space for (σ, F) separately. Lower bounds for the prior mass in appropriate neighbourhoods of the point (σ_0, F_0) are incorporated immediately.

Theorem 4.4. *Suppose that the regression family $\mathscr{F}$ is one of those specified in the beginning of section 4.2. Assume that the prior Π on $\mathscr{P}$ is of the product form indicated above. Then there exist constants $K, c, C > 0$ such that:*

$$\Pi\Big(B\big(K\delta\log(1/\delta); P_0\big)\Big) \geq C\exp\Big(-c\big(\log(1/\delta)\big)^3\Big)\,\Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\big),$$

for small enough δ.

Proof If the class of regression functions $\mathscr{F}$ is a Lipschitz-class with exponent in $(0,1)$, we set α equal to that exponent. In other cases we set $\alpha = 1$.

Let $\epsilon > 0$ be given. By lemma 4.9 there exists a discrete F_0' in D with at most $N_\epsilon = C(\log(1/\epsilon))^2$ support points $z_1, \ldots, z_{N_\epsilon}$ of the form $F_0' = \sum_{i=1}^{N_\epsilon} p_i\delta_{z_i}$ with $\sum_{i=1}^{N_\epsilon} p_i = 1$, such that:

$$\|p_{\sigma_0,f_0,F_0'} - p_{\sigma_0,f_0,F_0}\|_{1,\mu} \leq C'\epsilon^\alpha,$$

for some constant $C' > 0$. Although the assertion of lemma 4.9 is stronger, we include the power of α because we assume (without loss of generality) that the set of support points for F_0' is 2ϵ-separated. If this is not the case, take a maximal 2ϵ-separated subset and shift the masses of other support points of F_0' to points in the chosen subset within distance 2ϵ, to obtain a new discrete distribution F_0''. Arguing as in the proof of theorem 4.3, we see that the corresponding change in $L_1(\mu)$-distance between $p_{\sigma_0,f_0,F_0'}$ and $p_{\sigma_0,f_0,F_0''}$ is upper-bounded by a multiple of ϵ^α, since the family of regression functions satisfies (4.7) by assumption. The distribution function F_0'' so obtained may then replace F_0'. By lemma 4.11, there exists a constant $K_3 > 0$ such that for all $F \in D$:

$$\|p_{\sigma_0,f_0,F} - p_{\sigma_0,f_0,F'}\|_{1,\mu} \leq K_3\Big(\epsilon^\alpha + \sum_{i=1}^{N_\epsilon}\big|F[z_i - \epsilon, z_i + \epsilon] - p_i\big|\Big).$$

Let (σ, f, F) be a point in the parameter space of the model. The Hellinger distance between $p_{\sigma,f,F}$ and p_{σ_0,f_0,F_0} is upper-bounded as follows (for constants $K_1, K_2 > 0$):

$$\begin{aligned}
H(P_{\sigma,f,F},P_{\sigma_0,f_0,F_0}) &\leq H(P_{\sigma,f,F}, P_{\sigma_0,f,F}) + H(P_{\sigma_0,f,F}, P_{\sigma_0,f_0,F}) + H(P_{\sigma_0,f_0,F}, P_{\sigma_0,f_0,F_0}) \\
&\leq \big\|p_{\sigma,f,F} - p_{\sigma_0,f,F}\big\|_{1,\mu}^{1/2} + H(P_{\sigma_0,f,F}, P_{\sigma_0,f_0,F}) + \big\|p_{\sigma_0,f_0,F} - p_{\sigma_0,f_0,F_0}\big\|_{1,\mu}^{1/2} \\
&\leq K_1\big|\sigma - \sigma_0\big|^{1/2} + K_2\big\|f - f_0\big\| \\
&\quad + \Big(\big\|p_{\sigma_0,f_0,F} - p_{\sigma_0,f_0,F_0'}\big\|_{1,\mu} + \big\|p_{\sigma_0,f_0,F_0'} - p_{\sigma_0,f_0,F_0}\big\|_{1,\mu}\Big)^{1/2},
\end{aligned}$$

where we have used lemmas 4.5, 4.6 and corollary 4.1. Moreover, we see that there exists a constant $K_4 > 0$ such that for small enough $\eta > 0$ and $P \in \mathscr{P}$ such that $H(P, P_0) \leq \eta$:

$$-P_0\log\frac{p}{p_0} \vee P_0\Big(\log\frac{p}{p_0}\Big)^2 \leq K_4^2\eta^2\Big(\log\frac{1}{\eta}\Big)^2,$$

as a result of lemma 4.12. Combining the last two displays and using definition (4.3), we find that, for some constants $K_5, K_6 > 0$, the following inclusions hold:

$$\begin{aligned}\Big\{(\sigma, f, F) \in I \times \mathscr{F} \times D : |\sigma - \sigma_0|^{1/2} \leq \epsilon^{\alpha},\, \|f - f_0\| \leq \epsilon^{\alpha/2},\, \sum_{j=1}^{N_\epsilon} \big|F[z_j - \epsilon, z_j + \epsilon] - p_j\big| \leq \epsilon^{\alpha}\Big\} \\ \subset \Big\{(\sigma, f, F) \in I \times \mathscr{F} \times D : H(P_{\sigma,f,F}, P_0) \leq K_5 \epsilon^{\alpha/2}\Big\} \\ \Big\{P \in \mathscr{P} : H(P, P_0) \leq K_5 \epsilon^{\alpha/2}\Big\} \subset B\big(K_6 \epsilon^{\alpha/2} \log(1/\epsilon); P_0\big),\end{aligned} \tag{4.26}$$

for small enough ϵ and with the notation p_0 for the density of P_0 ($p_0 = p_{\sigma_0, f_0, F_0}$)). Using the fact that the prior measure of the rectangle set on the *l.h.s.* of the first inclusion above factorizes, we find that:

$$\begin{aligned}\Pi\Big(B\big(K_6 \epsilon^{\alpha/2} \log(1/\epsilon); P_0\big)\Big) \geq \Pi_I\big(\sigma \in I : |\sigma - \sigma_0|^{1/2} \leq \epsilon^{\alpha}\big)\, \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \epsilon^{\alpha/2}\big) \\ \times \Pi_D\Big(F \in D : \sum_{j=1}^{N_\epsilon} \big|F[z_j - \epsilon, z_j + \epsilon] - p_j\big| \leq \epsilon^{\alpha}\Big).\end{aligned}$$

Note that $\epsilon^{\alpha} \geq \epsilon$ for small enough ϵ, so that

$$\Pi_D\Big(\sum_{j=1}^{N_\epsilon} \big|F[z_j - \epsilon, z_j + \epsilon] - p_j\big| \leq \epsilon^{\alpha}\Big) \geq \Pi_D\Big(\sum_{j=1}^{N_\epsilon} \big|F[z_j - \epsilon, z_j + \epsilon] - p_j\big| \leq \epsilon\Big).$$

According to lemma 6.1 in Ghosal *et al.* (2000) [39] (also given as lemma A.2 in Ghosal and Van der Vaart (2001) [40]), there are constants $C', c' > 0$ such that

$$\Pi_D\Big(\sum_{j=1}^{N_\epsilon} \big|F[z_j - \epsilon, z_j + \epsilon] - p_j\big| \leq \epsilon\Big) \geq C' \exp\big(-c' N_\epsilon \log(1/\epsilon)\big) \geq C' \exp\big(-c' C \big(\log(1/\epsilon)\big)^3\big).$$

Furthermore, continuity and strict positivity of the density of the prior Π_I imply that (see the proof of lemma 4.17):

$$\Pi_I(\sigma \in I : |\sigma - \sigma_0| \leq \epsilon^{\alpha}) \geq \pi_1\, \epsilon^{\alpha} = \pi_1 \exp\big(-\alpha \log(1/\epsilon)\big),$$

for some constant $\pi_1 > 0$. Note that the exponent on the *r.h.s.* falls above all multiples of $-(\log(1/\epsilon))^3$ for small enough ϵ. Substitution of $\delta = \epsilon^{\alpha/2}$ leads to the conclusion that there exist constants $K, c, C > 0$ such that:

$$\Pi\Big(B\big(K\delta \log(1/\delta); P_0\big)\Big) \geq C \exp\Big(-c\big(\log(1/\delta)\big)^3\Big)\, \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\big),$$

for small enough δ. □

If the model is not identifiable in the parameter space $I \times \mathscr{F} \times D$, the above conditions are more stringent than necessary. The point (σ_0, f_0, F_0) may not be the only one that is mapped to P_0, so the first inclusion in (4.26) may discount parts of the parameter space that also contribute to the Kullback-Leibler neighbourhoods $B(\epsilon; P_0)$. However, the methods we use to lower-bound the prior mass rely on uniformity in the sense that neighbourhoods of *every* point in the parameter space receive a certain minimal fraction of the total prior mass. Therefore, identifiability issues do not affect the argument.

4.4.1 Lemmas

In the following lemma, it is assumed that the regression class $\mathscr{F}$ is one of those specified in the beginning of section 4.2. If the class of regression functions is a Lipschitz-class with exponent in $(0,1)$, we set α equal to that exponent. In other cases we set $\alpha = 1$.

Lemma 4.11. *Let $\epsilon > 0$ be given and let $F' = \sum_{i=1}^N p_i \delta_{z_i}$ be a convex combination of point-masses, where the set $\{z_i : i = 1, \ldots, N\}$ is 2ϵ-separated. Then there exists a constant $K > 0$ such that for all $\sigma \in I$, $f \in \mathscr{F}$ and all $F \in D$:*

$$\|p_{\sigma,f,F} - p_{\sigma,f,F'}\|_{1,\mu} \leq K\Big(\epsilon^\alpha + \sum_{i=1}^N \big|F[z_i - \epsilon, z_i + \epsilon] - p_i\big|\Big),$$

for small enough ϵ.

Proof Let F be given. We partition the real line by $\mathbb{R} = \cup_i A_i \cup B$, with $B = \big(\cap_i B_i\big)$, where

$$A_i = \big\{z : |z - z_i| \leq \epsilon\big\}, \quad B_i = \big\{z : |z - z_i| > \epsilon\big\},$$

and decompose the absolute difference between $p_{\sigma,f,F}$ and $p_{\sigma,f,F'}$ accordingly:

$$\begin{aligned}\big|p_{\sigma,f,F} - p_{\sigma,f,F'}\big|(x,y) &= \Big|\int_{\mathbb{R}} \varphi_\sigma(x - z)\varphi_\sigma\big(y - f(z)\big)\, d(F - F')(z)\Big| \\ &= \Big|\sum_{i=1}^N \int_{A_i} \varphi_\sigma(x - z)\varphi_\sigma(y - f(z))\, d(F - F')(z) + \int_B \varphi_\sigma(x - z)\varphi_\sigma(y - f(z))\, dF(z)\Big|,\end{aligned}$$

for all $(x,y) \in \mathbb{R}^2$. Integrating this expression over $\mathbb{R}^2$, we find that the $L_1(\mu)$-difference is bounded as follows:

$$\begin{aligned}\big\|p_{\sigma,f,F} - p_{\sigma,f,F'}\big\|_{1,\mu} &\leq \sum_{i=1}^N \big|F[z_i - \epsilon, z_i + \epsilon] - p_i\big| + F\Big(\bigcap_{i=1}^N B_i\Big) \\ &\quad + \sum_{i=1}^N \int_{A_i}\int_{\mathbb{R}^2} \big|\varphi_\sigma(x - z)\varphi_\sigma(y - f(z)) - \varphi_\sigma(x - z_i)\varphi_\sigma(y - f(z_i))\big|\, d\mu(x,y)\, dF(z),\end{aligned}$$

by Fubini's theorem and the triangle inequality. To upper-bound the last term on the *r.h.s.* in the above display, we use that for all $x, y \in \mathbb{R}$ and $z \in [-A, A]$:

$$\begin{aligned}&\big|\varphi_\sigma(x - z)\varphi_\sigma(y - f(z)) - \varphi_\sigma(x - z_i)\varphi_\sigma(y - f(z_i))\big| \\ &\quad \leq \big|\varphi_\sigma(x - z) - \varphi_\sigma(x - z_i)\big|\varphi_\sigma(y - f(z)) + \big|\varphi_\sigma(y - f(z)) - \varphi_\sigma(y - f(z_i))\big|\varphi_\sigma(x - z_i),\end{aligned}$$

and argue as in the proof of lemma 4.7, to see that the integrand is bounded by a multiple of $|z - z_i|^\alpha$ for small enough ϵ. Noting that the intervals $[z_i - \epsilon, z_i + \epsilon]$ are disjoint due to 2ϵ-separation of the set $\{z_i : i = 1, \ldots, N\}$, we see that there exists a constant $L' > 0$ such that

$$\big\|p_{\sigma,f,F} - p_{\sigma,f,F'}\big\|_{1,\mu} \leq L'\epsilon^\alpha + \sum_{i=1}^N \big|F[z_i - \epsilon, z_i + \epsilon] - p_i\big| + F\Big(\bigcap_{i=1}^N B_i\Big).$$

Furthermore, by De Morgan's law and the disjointness of the intervals $[z_i - \epsilon, z_i + \epsilon]$:

$$F\Big(\bigcap_{i=1}^{N}\{z : |z - z_i| > \epsilon\}\Big) = 1 - F\Big(\bigcup_{i=1}^{N}\{z : |z - z_i| \leq \epsilon\}\Big)$$
$$= \sum_{i=1}^{N} p_i - \sum_{i=1}^{N} F[z_i - \epsilon, z_i + \epsilon] \leq \sum_{i=1}^{N} \big|F[z_i - \epsilon, z_i + \epsilon] - p_i\big|,$$

which proves the assertion. □

Lemma 4.12. *Let $P, Q \in \mathscr{P}$ be given. There exists a constant $K > 0$ such that for small enough $H(P, Q)$:*

$$\int p \log \frac{p}{q}\, d\mu \leq K^2 H^2(P, Q)\Big(\log \frac{1}{H(P, Q)}\Big)^2,$$
$$\int p\Big(\log \frac{p}{q}\Big)^2 d\mu \leq K^2 H^2(P, Q)\Big(\log \frac{1}{H(P, Q)}\Big)^2. \tag{4.27}$$

The constant K does not depend on P, Q.

Proof Fix $\delta \in (0, 1]$ and consider the integral:

$$M_\delta^2 = \int p\Big(\frac{p}{q}\Big)^\delta d\mu.$$

We shall prove that for a suitable choice of δ, $M_\delta^2 < \infty$. Since all densities involved are bounded away from zero and infinity on compacta, we consider only the domain $O = \mathbb{R}^2 \setminus [-C, C] \times [-C, C]$, for some large constant $C \geq A \vee B$. Note that:

$$\int_O p\Big(\frac{p}{q}\Big)^\delta d\mu \leq \int_O U\Big(\frac{U}{L}\Big)^\delta d\mu,$$

where (L, U) forms an envelope for the model. This envelope follows from the fact that the regression densities (4.2) fall in the class of mixture densities obtained by mixing the normal kernel $\varphi_\sigma(x)\varphi_\sigma(y)$ on $\mathbb{R}^2$ by means of a two-dimensional distribution that places all its mass in the rectangle $[-A, A] \times [-B, B]$. There exists a lower bound for this envelope which factorizes into x- and y-envelopes (L_X, U_X) and (L_Y, U_Y) that are constant on sets that include $[-A, A]$ and $[-B, B]$ respectively and have Gaussian tails. The domain O can therefore be partitioned into four subdomains in which either x or y is bounded and four subdomains in which both coordinates are unbounded. Reflection-symmetries of the envelope functions suffice to demonstrate that integrals of $U(U/L)^\delta$ can be expressed as products of trivially finite factors and integrals of the form:

$$\int_L^\infty U_X(x)\Big(\frac{U_X}{L_X}\Big)^\delta(x)\, d\mu(x), \quad \int_L^\infty U_Y(y)\Big(\frac{U_Y}{L_Y}\Big)^\delta(y)\, d\mu(y).$$

For large enough C, the envelope functions $L_X(x)$ and $U_X(x)$) are equal to multiples of $\varphi_{\underline{\sigma}}(x + A)$ and $\varphi_{\overline{\sigma}}(x - A)$) on the domain (C, ∞) and hence, for some constants $c, K > 0$:

$$\int_L^\infty U_X(x)\Big(\frac{U_X}{L_X}\Big)^\delta(x)\, d\mu(x) \leq K \int_L^\infty e^{c\delta x^2}\, \varphi_{\overline{\sigma}}(x - A)\, dx,$$

which is finite for small enough $\delta > 0$. Similarly, one can prove finiteness of the integrals over y. This proves that the condition for theorem 5 in Wong and Shen (1995) [97] is satisfied. Note that the choice for δ is independent of p, q. Furthermore, the value of M_δ can be upper-bounded independent of p, q, as is apparent from the above. Hence, for small enough $\eta > 0$, (4.27) holds. □

4.5 Regression classes

Theorems 4.3 and 4.4 demonstrate that both the entropy and prior-mass conditions in theorem 4.1 can be decomposed in a term that pertains to the regression function f and a term pertaining to the parameters (σ, F). This makes it possible to consider entropy and prior-mass restricted to the regression class separately.

In the first subsection, we state a bound on the metric entropy of the classes $C_{\beta,M}[-A, A]$ due to Kolmogorov, who derived it shortly after his introduction of the concept of covering numbers. This bound is used in the second subsection to demonstrate that so-called *net priors* can be used for non-parametric regression classes in this situation. Also discussed is an alternative approach, that uses (adapted versions of) Jackson's approximation theorem. Up to a logarithmic correction, the second approach reproduces Kolmogorov's bound for the metric entropy, but upon application in the form of so-called *sieve-priors*, the resulting lower bounds for the prior mass in neighbourhoods of the true regression function are sub-optimal in a more grave manner. Nevertheless, we indulge in an explanation of the second approach, because it provides a good example of the methods and subtleties of Bayesian procedures in non-parametric problems. We also give the necessary bounds on the entropy and prior mass of parametric regression classes.

4.5.1 Covering numbers of regression classes

The usefulness of bounds (4.17) and (4.18) indicates that the class of regression functions parametrizing the model is best chosen within the (Banach-)space $C[-A, A]$ of continuous functions on the closed interval $[-A, A]$ with the uniform norm $\|\,.\,\|$. According to the Weierstrass approximation, polynomials are dense in $C[-A, A]$; bounded families of polynomials can therefore be used to approximate regression families $\mathscr{F}$ as characterised in point *(c)* at the beginning of subsection 4.1.1. The Ascoli-Arzelà theorem asserts that if, in addition, $\mathscr{F}$ is equicontinuous, it is relatively compact. Hence bounded, equicontinuous families $\mathscr{F}$ are totally bounded in the norm-topology, rendering covering numbers finite,

$$N(\epsilon, \mathscr{F}, \|\,.\,\|) < \infty, \tag{4.28}$$

for all $\epsilon > 0$. As a side-remark, note that Schwartz' conditions for consistency of Bayesian procedures (see Schwarz (1965) [82]) require the existence of an asymptotically consistent

sequence of test functions, which can be inferred from finiteness of covering numbers like above (see, for instance, the subsection on consistency in [57]).

However, since we are interested in rates of convergence, finiteness of covering numbers is not enough and a more detailed analysis of the behaviour of $N(\epsilon, \mathscr{F}, \|\,.\,\|)$ for small ϵ is needed. We reproduce here a result due to Kolmogorov and Tikhomirov (1961) [60] (in a version as presented in Van der Vaart and Wellner (1996) [89]), that gives the required bound:

Lemma 4.13. *Let $\beta > 0$, $M > 0$ be given. There exists a constant K depending only on β and A, such that:*

$$\log N\big(\epsilon, C_{\beta,M}[-A, A], \|\cdot\|\big) \leq K\Big(\frac{1}{\epsilon}\Big)^{1/\beta}, \tag{4.29}$$

for all $\epsilon > 0$.

The proof of this lemma is a special version of the proof of theorem 2.7.1 in [89], which consists of a fairly technical approximation by polynomials. To improve our understanding of the above result, we briefly digress on an approach that is based on Jackson's approximation theorem.

Fix an $n \geq 1$; Jackson's approximation theorem (see Jackson (1930) [48]) says that if $f \in \mathrm{Lip}_M(\alpha)$, there exists an n-th order polynomial p_n such that:

$$\|f - p_n\| \leq \frac{K}{n^\alpha}, \tag{4.30}$$

where $K > 0$ is a constant that depends only on A and M. Moreover, if $f \in D_{\alpha,M}(q)$, there exists a polynomial p_n of degree n such that:

$$\|f - p_n\| \leq \frac{K'}{n^{q+\alpha}}, \tag{4.31}$$

where $K' > 0$ is a constant that depends on A, q, α and M. Indeed, in its most general formulation, Jackson's theorem applies to arbitrary continuous functions f, relating the degree of approximation to the modulus of continuity. As such, it provides a more precise version of Weierstrass' theorem.

The class of *all* n-th degree polynomials is larger than needed for the purpose of defining nets over the bounded regression classes we are interested in. Let $B > 0$ denote the constant that bounds all functions in $\mathscr{F}$. With given $\gamma > 0$, define $P_n' = \{p \in P_n : \|p\| \leq (1+\gamma)B\}$. By virtue of the triangle inequality, any polynomial used to approximate f as in (4.30) or (4.31) satisfies a bound slightly above and arbitrarily close to B with increasing n. Hence, for large enough n, P_n' is a L/n^β-net over $C_{\beta,M}[-A, A]$, where $L > 0$ is a constant that depends only on the constants defining the regression class. For these finite-dimensional, bounded subsets of $C[-A, A]$, the order of suitable nets can be calculated. The upper-bound for the metric entropy of Lipschitz and smoothness classes based on Jackson's theorem takes the following form.

Lemma 4.14. *Let $\beta > 0$ and $M > 0$ be given. There exists a constant $K' > 0$ such that:*

$$\log N(\epsilon, C_{\beta,M}[-A, A], \|\,.\,\|) \leq K'\Big(\frac{1}{\epsilon}\Big)^{1/\beta} \log \frac{1}{\epsilon},$$

for small enough $\epsilon > 0$.

Proof Let $\epsilon > 0$ be given and choose n to be the smallest integer satisfying $n^\beta \geq 1/\epsilon$. Define $P_n'' = \{p \in P_n : \|p\| \leq L\}$ for some $L > B$. As argued after (4.31), there is a uniformly bounded set P_n' of polynomials of degree n that forms an ϵ-net over $C_{\beta,M}[-A, A]$. If n is chosen large enough, P_n' is a proper subset of P_n''. To calculate an upper bound for the covering number of P_n', let $\delta > 0$ be given and let $p_1, \ldots, p_D$ be a (maximal) set of δ-separated polynomials in P_n', where D is the packing number $D(\delta, P_n', \|\,.\,\|)$. Note that the balls $B_i = \{p \in P_n' : \|p - p_i\| < \frac{1}{2}\delta\}$, $(i = 1, \ldots, D)$, do not intersect. If δ is chosen small enough, $B_i \subset P_n''$. The linear map $\hat{p} : \mathbb{R}^{n+1} \to P_n$ that takes a vector $(a_0, ..., a_n)$ into the polynomial $\sum_{m=0}^n a_m z^m$ is Borel measurable and is used to define the sets $C_i = \hat{p}^{-1}(B_i)$. Note that the sets C_i are obtained from $C = \hat{p}^{-1}(P_n'')$ by rescaling and translation for all i. By the same argument as used in the proof of lemma 4.32, we conclude that there is a constant L such that the packing number satisfies:

$$D(\delta, P_n', \|\,.\,\|) \leq \Big(\frac{L}{\delta}\Big)^{n+1},$$

for small enough $\delta > 0$, which serves as an upper bound for the covering number as well. Choosing δ equal to a suitable multiple of $n^{-\beta}$ for large enough n, we find a constant $K' > 0$ and a net over $C_{\beta,M}[-A, A]$ in P_n of order bounded by $(K'n^\beta)^{n+1}$. The triangle inequality then guarantees the existence of a slightly less dense net over $C_{\beta,M}[-A, A]$ inside $C_{\beta,M}[-A, A]$ of the same order. We conclude that there exists a constant $K'' > 0$ such that:

$$\log N(\epsilon, C_{\beta,M}[-A, A], \|\,.\,\|) \leq K'' n \log n^\beta,$$

for large enough n, which leads to the stated bound upon substitution of the relation between ϵ and n. □

The power of ϵ in the bound asserted by the above lemma is that of lemma 4.13. The logarithmic correction can be traced back to the n-dependence of the radius of the covering balls B_i, *i.e.* the necessity of using finer and finer nets over P_n' to match the n-dependence in the degree of approximation. Therefore, there is no obvious way of adapting the above proof to eliminate the $\log(1/\epsilon)$-factor and Kolmogorov's approach gives a strictly smaller bound on the entropy. However, the above illustrates the origin of the β-dependence in the power of ϵ more clearly.

For parametric classes (as given under *(iii)* in the beginning of section 4.2), the entropy is bounded in the following lemma.

Lemma 4.15. *For a parametric class $\mathscr{F}_\Theta$, there exists a constant $K > 0$ such that the metric entropy is bounded as follows:*

$$\log N(\epsilon, \mathscr{F}_\Theta, \|\,.\,\|) \leq K \log \frac{1}{\epsilon}, \tag{4.32}$$

for small enough $\epsilon > 0$.

Proof Since, by assumption, $\Theta \subset \mathbb{R}^k$ is bounded by some constant $M' > 0$, the covering numbers of Θ are upper-bounded by the covering numbers of the ball $B(M', 0) \subset \mathbb{R}^k$ of radius M' centred on 0. Let $\delta > 0$ be given. Since covering numbers are bounded by packing numbers, we see that:

$$N(\delta, \Theta, \| \, . \, \|_{\mathbb{R}^k}) \leq D(\delta, B(M', 0), \| \, . \, \|_{\mathbb{R}^k}).$$

Let $\theta_1, \ldots, \theta_D$ (with $D = D(\delta, B(M', 0), \| \, . \, \|_{\mathbb{R}^k})$) be a maximal δ-separated subset of $B(M', 0)$. The balls $B_i = B(\frac{1}{2}\delta, \theta_i)$ do not intersect and are all contained in the ball $B(M' + \frac{1}{2}\delta, 0)$ by virtue of the triangle inequality. Therefore, the sum of the volumes of the balls B_i (which are all equal and proportional to $(\frac{1}{2}\delta)^k$, due to translation invariance and scaling behaviour of the Lebesgue measure) lies below the volume of the ball $B(M' + \frac{1}{2}\delta, 0)$. We conclude that:

$$D(\delta, B(M', 0), \| \, . \, \|_{\mathbb{R}^k})(\tfrac{1}{2}\delta)^k \leq (M' + \tfrac{1}{2}\delta)^k.$$

Assuming that $\delta < 2M'$, we see that:

$$D(\delta, B(M', 0), \| \, . \, \|_{\mathbb{R}^k}) \leq \Big(\frac{4M'}{\delta}\Big)^k. \tag{4.33}$$

Next, note that due to (4.8), any δ-net over Θ leads to a $L\delta^\rho$-net over the regression class $\mathscr{F}_\Theta$, whence we see that:

$$N(L\delta^\rho, \mathscr{F}_\Theta, \| \, . \, \|) \leq N(\delta, \Theta, \| \, . \, \|_{\mathbb{R}^k}). \tag{4.34}$$

Let $\epsilon > 0$ be given and choose $\delta = (\epsilon/L)^{1/\rho}$. Combining (4.33) and (4.34), we find that there exists a constant $K > 0$ such that:

$$\log N(\epsilon, \mathscr{F}_\Theta, \| \, . \, \|) \leq K \log \frac{1}{\epsilon},$$

for small enough ϵ. □

These bounds on the small-ϵ behaviour of the entropy are incorporated in the calculation of bounds for the entropy of the errors-in-variables model through theorem 4.3.

4.5.2 Priors on regression classes

This subsection is devoted to the definition of a suitable prior $\Pi_{\mathscr{F}}$ on the regression class $\mathscr{F}$. The challenge is to show that $\Pi_{\mathscr{F}}$ places 'enough' mass in small neighbourhoods of any point in the regression class. More specifically, a lower bound is needed for the prior mass of neighbourhoods of the (unknown) regression function $f_0 \in \mathscr{F}$:

$$\Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\big), \tag{4.35}$$

for small enough $\delta > 0$ (refer to theorem 4.4).

Jackson's theorem suggests that a natural definition of a prior on $\mathscr{F}$ entails the placement of prior mass on all (finite-dimensional) linear spaces of n-th degree polynomials P_n on $[-A, A]$,

since their union is dense in $C[-A, A]$ and therefore also in $\mathscr{F}$. Fix the regression class $\mathscr{F}$. For all $n \geq 1$ we define:

$$\mathscr{F}_n = \mathscr{F} \cap P_n,$$

i.e. the subsets of n-th degree polynomials in the regression class[2]. Note that $\mathscr{F}_n \subset \mathscr{F}_{n+1}$ for all n, and that $\mathscr{F}$ lies in the closure of their union. The linear map $\hat{p} : \mathbb{R}^{n+1} \to P_n$ that takes a vector $(a_0, ..., a_n)$ into the polynomial $\sum_{m=0}^{n} a_m z^m$ can be used to define a subset $\hat{p}^{-1}(\mathscr{F}_n) \subset \mathbb{R}^k$ with Lebesgue measure strictly above zero. Normalizing the Lebesgue measure to 1 on $\hat{p}^{-1}(\mathscr{F}_n)$, the inverse map $\hat{p}^{-1}$ serves to define a probability measure Π_n on $\mathscr{F}_n$. Any sequence $(b_n)_{n\geq 0}$ such that $b_n \geq 0$ and $\sum_{n=0}^{\infty} b_n = 1$, may be used to define a prior $\Pi_{\mathscr{F}}$ by the infinite convex combination:

$$\Pi_{\mathscr{F}}(A) = \sum_{n=0}^{\infty} b_n \Pi_n(A) = \sum_{n=0}^{\infty} b_n \Pi_n(A \cap \mathscr{F}_n), \tag{4.36}$$

for all A in the Borel σ-algebra generated by the norm topology on $\mathscr{F}$. Following Huang [45], we refer to priors obtained in this manner as *sieve priors* (although arguably the sequence $(\Pi_n)_{n\geq 1}$ is more deserving of this name than the (n-independent) prior $\Pi_{\mathscr{F}}$).

With a sieve prior, a proof of (4.4) amounts to showing that neighbourhoods of f_0 have intersections with the sets $\mathscr{F}_n$ and that the sum of the masses of these intersections is large enough. Obviously, Jackson's approximation provides a useful way to assert that balls centred on f_0 intersect with all P_n' from a certain minimal n onward. However, as is apparent from (4.35), this is not sufficient, because the relevant neighbourhoods are restricted to the regression class $\mathscr{F}$. One would have to show that these *restricted* neighbourhoods intersect with the sets $\mathscr{F}_n$.

Jackson's theorem does not assert anything concerning Lipschitz-bounds of the approximating polynomial or derivatives thereof. The assertion that p_n approximates f in uniform norm leaves room for very sharp fluctuations of p_n on small scales, even though it stays within a bracket of the form $[f - K/n^{\beta}, f + K/n^{\beta}]$. It is therefore possible that p_n lies far outside $\mathscr{F}_n$, rendering neighbourhoods of p_n in P_n unfit for the purpose. Although it is possible to adapt Jackson's theorem in such a way that the approximating polynomials satisfy a Lipschitz condition that is arbitrarily close to that of the regression class, this adaptation comes at a price with regard to the degree of approximation. As it turns out, this price leads to substantial corrections for the rate of convergence and ultimately to sub-optimality (with respect to the *power* of ϵ rather than logarithmically). That is not to say that sieve priors are in any sense sub-optimal. (Indeed, sieve priors have been used with considerable success in certain situations; for an interesting example, see the developments in adaptive Bayesian estimation, for

[2] Alternatively, intersection of the spaces P_n with the model could be omitted and prior mass could be placed on the entire space P_n (for every $n \geq 1$). The support of the resulting prior Π_F (*c.f.* definition (4.36)) is then strictly larger than the model $\mathscr{F}$ and theorem 4.1 no longer applies in the form given. Indeed, placing prior mass on the approximating spaces rather than the model proper, would introduce new problems that manifest themselves through the entropy condition (4.5) rather than the prior mass condition (4.4).

instance in Huang [45].) The calculation underlying the claims made above merely shows that the construction via adapted versions of Jackson's theorem does not lead to optimal results, leaving the possibility that a sieve prior satisfies (4.4) open. What it does show, however, is that this may be very hard to demonstrate.

Therefore, we define the prior on the regression class in a different fashion, based on the upper bounds for covering numbers obtained in the previous subsection. Let the regression class $\mathscr{F}$ be a bounded, equicontinuous subset of $C[-A, A]$, so that the covering numbers $N(\epsilon, \mathscr{F}, \|\,.\,\|)$ are finite for all $\epsilon > 0$. Let $(a_m)_{m\geq 1}$ be a monotonically decreasing sequence, satisfying $a_m > 0$ (for all $m \geq 1$), and $a_m \downarrow 0$. For every $m \geq 1$, there exists an a_m-net $\{f_i \in \mathscr{F} : i = 1, \ldots, N_m\}$ over $\mathscr{F}$, where $N_m = N(a_m, \mathscr{F}, \|\,.\,\|)$. We define, for every $m \geq 1$, a discrete probability measure Π_m that distributes its mass uniformly over the set $\{f_i : i = 1, \ldots, N_m\}$:

$$\Pi_m = \sum_{i=1}^{N_m} \frac{1}{N_m} \delta_{f_i}.$$

Any sequence $(b_n)_{n\geq 0}$ such that $b_n \geq 0$ and $\sum_{n=0}^{\infty} b_n = 1$, may be used to define a prior $\Pi_{\mathscr{F}}$ on $\mathscr{F}$ by the infinite convex combination:

$$\Pi_{\mathscr{F}}(A) = \sum_{m=0}^{\infty} b_m \Pi_m(A), \tag{4.37}$$

for all A in the Borel σ-algebra generated by the norm topology on $\mathscr{F}$. Priors defined in this manner are referred to as a *net priors* and resemble those defined in Ghosal, Ghosh and Ramamoorthi (1997) [38], (see also, Ghosal *et al.* (2000) [39]).

Note that for all $m \geq 1$ and every $f \in \mathscr{F}$, there is an f_i satisfying $\|f - f_i\| \leq a_m$. So for every $f_0 \in \mathscr{F}$ and all $\delta > 0$, we have:

$$\Pi_m\big(\|f - f_0\| \leq \delta\big) \geq \frac{1}{N_m},$$

if $a_m \leq \delta$, *i.e.* for all m large enough. This means that the priors Π_m satisfy lower bounds for the mass in neighbourhoods of points in the regression class, that are inversely related to upper bounds satisfied by the covering numbers. As is demonstrated below, choices for the sequences a_m and b_m exist such that this property carries over to a prior of the form (4.37).

Lemma 4.16. *Let $\beta > 0$ and $M > 0$ be given and define $\mathscr{F}$ to be the class $C_{\beta,M}[-A, A]$. There exists a net prior $\Pi_{\mathscr{F}}$ and a constant $K > 0$ such that*

$$\log \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\big) \geq -K \frac{1}{\delta^{1/\beta}}, \tag{4.38}$$

for small enough δ.

Proof Define, for all $m \geq 1$, $a_m = m^{-\beta}$. Then the covering number N_m satisfies, for some constant $K' > 0$:

$$\log N_m = \log N(a_m, \mathscr{F}, \|\,.\,\|) \leq K' a_m^{-1/\beta} = K' m,$$

according to lemma 4.13. Let $\delta > 0$ be given and choose the sequence $b_m = (1/2)^m$. Let M be an integer such that:

$$\frac{1}{\delta^{1/\beta}} \leq M \leq \frac{1}{\delta^{1/\beta}} + 1.$$

Then for all $m \geq M$, $a_m \leq \delta$ and, due to the inequality (4.37), the net prior $\Pi_{\mathscr{F}}$ satisfies:

$$\begin{aligned} \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\big) &\geq \sum_{m \geq M} b_m \Pi_m\big(\|f - f_0\| \leq \delta\big) \geq \sum_{m \geq M} \Big(\frac{e^{-K'}}{2}\Big)^m \\ &\geq \tfrac{1}{2} e^{-K'M} \geq \tfrac{1}{2} e^{-K'(\delta^{-1/\beta}+1)} \geq \tfrac{1}{2} e^{-2K'\delta^{-1/\beta}}, \end{aligned} \tag{4.39}$$

for small enough δ. □

For parametric classes, the prior mass in neighbourhoods of f_0 is lower-bounded in the following lemma.

Lemma 4.17. *Assume that the regression class $\mathscr{F}$ is parametric: $\mathscr{F} = \mathscr{F}_\Theta$. Any prior Π_Θ on Θ induces a prior $\Pi_{\mathscr{F}}$ with the Borel σ-algebra generated by the topology of the norm $\|\,.\,\|$ as its domain. Furthermore, if Π_Θ is dominated by the Lebesgue measure and has a density that is strictly positive at θ_0, then there exists a constant $R > 0$ such that the prior mass in neighbourhoods of f_0 is bounded as follows:*

$$\log \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \epsilon\big) \geq -R \log \frac{1}{\epsilon}, \tag{4.40}$$

for small enough $\epsilon > 0$.

Proof The Lipschitz condition (4.8) ensures that the map $\hat{f} : \Theta \to \mathscr{F}_\Theta : \theta \mapsto f_\theta$ is continuous, implying measurability with respect to the corresponding Borel σ-algebras. So composition of Π_Θ with $\hat{f}^{-1}$ induces a suitable prior on $\mathscr{F}_\Theta$. As for the second assertion, let $\delta > 0$ be given. Since Π_Θ has a continuous Lebesgue density $\pi : \Theta \to \mathbb{R}$ that satisfies $\pi(\theta_0) > 0$ by assumption and since θ_0 is internal to Θ, there exists an open neighbourhood $U \subset \Theta$ of θ_0 and a constant $\pi_1 > 0$ such that $\pi(\theta) \geq \pi_1$ for all $\theta \in U$. Therefore, for all balls $B(\delta, \theta_0) \subset U$ (*i.e.* for small enough $\delta > 0$), we have:

$$\Pi_\Theta\big(B(\delta, \theta_0)\big) = \int_{B(\delta,\theta_0)} \pi(\theta)\, d\theta \geq V_k \pi_1 \delta^k,$$

where V_k is the Lebesgue measure of the unit ball in $\mathbb{R}^k$. Note that due to property (4.8),

$$\big\{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta\big\} \subset \big\{\theta \in \Theta : \|f_\theta - f_0\| \leq L\delta^\rho\big\},$$

so that, for given $\epsilon > 0$ and the choice $\delta = (\epsilon/L)^{1/\rho}$:

$$\begin{aligned} \log \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \epsilon\big) &\geq \log \Pi_\Theta\big(\theta \in \Theta : \|\theta - \theta_0\| \leq (\epsilon/L)^{1/\rho}\big) \\ &\geq \log\big(V_k \pi_1 (\epsilon/L)^{k/\rho}\big) \geq -R \log \frac{1}{\epsilon}, \end{aligned}$$

for some constant $R > 0$ and small enough ϵ. □

The bounds on the small-ϵ behaviour of prior mass presented in this subsection are incorporated in the calculation of bounds for the prior mass of Kullback-Leibler neighbourhoods $B(\epsilon; P_0)$ through theorem 4.4.

4.6 Optimal rate for the posterior of θ

Assume, from here on, that the family of regression functions is parametric, *i.e.* parametrized by an open subset Θ of $\mathbb{R}^k$ for some finite dimension $k \geq 1$. The most familiar example is the linear errors-in-variables model, but the regression families studied in Taupin (2001) [86] are also parametric. On the basis of the semi-parametric discussion, it is clear that point-estimators for θ, at least in special cases, achieve rate $1/\sqrt{n}$ and can even be efficient. Point-estimators derived from the posterior distribution given in previous sections cannot be shown to converge at $\sqrt{n}$-rate, because the rate of convergence for the posterior is strictly lower (see, *e.g.*, theorem 2.5 in Ghosal *et al.* (2000) [39]). The purpose of the present section is to understand under which conditions a posterior distribution for the parameter θ alone converges to θ_0 at rate $1/\sqrt{n}$. In part, this is motivated by the more complicated problem of deriving a Bernstein-Von-Mises theorem in semi-parametric estimation problems (see Shen (2002) [84]). It is stressed that this section does not constitute a rigorous treatment, and merely indicates future possibilities based on earlier results.

We assume that the random variables e_1 and e_2 are normally distributed with mean 0 and have a known variances, both equal to 1. Hence the model $\mathscr{P}$ consists of probability measures $P_{\theta,F} = P_{\sigma=1,f_\theta,F}$, defined as in (4.2). The model is still assumed well-specified, *i.e.* there exist $\theta_0 \in \Theta$ and $F_0 \in D$ such that $P_0 = P_{\theta_0,F_0}$. Although not essential to the arguments that follow, we also assume that the model is identifiable (which is reasonable for example in the linear errors-in-variables model (see the introduction of [13])). We introduce a metric on the space $\Theta \times D$ as induced by the L_1-metric for the densities $p_{\theta,F}$. The fact that the parameter (θ_0, F_0) can be estimated consistently (which follows from the rate result derived in earlier sections) allows us to restrict attention to neighbourhoods of the point of convergence. We assume θ_0 is an internal point of Θ and we assume that the restriction to Θ of the induced L_1-topology is equivalent to the Euclidean topology.

The Bayesian procedure for errors-in-variables estimation as given in earlier sections gives rise to posterior convergence in Hellinger distance at $1/\sqrt{n}$-rate corrected by a $(\log n)^{3/2}$ factor. However, the parameter θ_0 can be (point-)estimated at rate $1/\sqrt{n}$, which suggests that the posterior converges non-isotropically, in the following sense: picture the non-parametric model in $\mathbb{R}^2$, where θ varies over the horizontal axis and F over the vertical (see figure 4.1).

The non-parametric rate $(\log n)^{3/2}/\sqrt{n}$ for the pair (θ, F) corresponds to the shrinking sequence of circles centred on the point of convergence P_0 and the results of previous sections show that this sequence captures all posterior mass asymptotically. This does not preclude the possibility that a sequence of smaller ellipses exists which also support all posterior mass asymptotically, and moreover, that the axes spanning these ellipses shrink at different rates. More particularly, it is possible that along the M-axis, the rate $(\log n)^{3/2}/\sqrt{n}$ has to be maintained, whereas the N-axis allows for the faster rate $1/\sqrt{n}$. In that case, a (skew) projection along the M-axis onto the θ-axis should suffice to achieve $1/\sqrt{n}$-rate.

In small neighbourhoods of P_0, the above phenomenon is governed by the behaviour of the likelihood. Continuing in a heuristic sense, one can think of the posterior density as proportional to the likelihood asymptotically in small neighbourhoods of the point of convergence. Asymptotic concentration of posterior mass in shrinking ellipses can then be inferred from a suitable (*i.e.* second order) expansion of the likelihood. The approach we use here allows us to restrict attention to the behaviour of the P_0-expectation of the log-likelihood, or equivalently, the Kullback-Leibler divergence with respect to P_0.

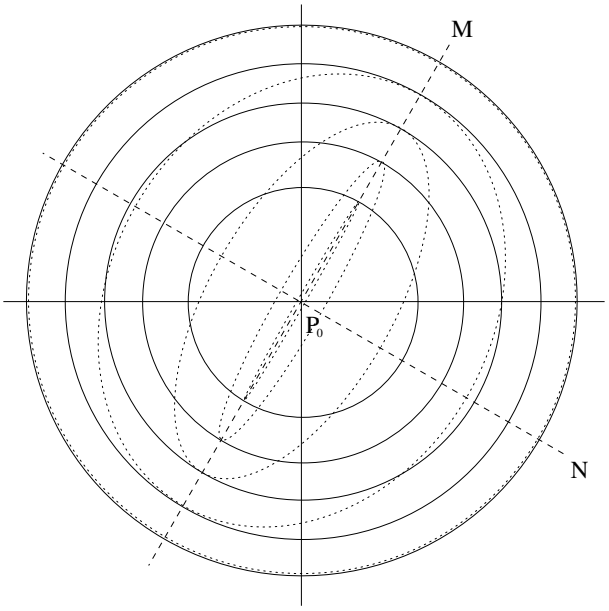

FIGURE 4.1 Convergence of the posterior distribution to P_0. The horizontal axis represents the parametric part $\theta \in \Theta$; the vertical axis the non-parametric $F \in D$. The sequence of concentric spheres are neighbourhoods of P_0 that support almost all mass of the posterior and decrease in radius at rate $(\log n)^{3/2}/\sqrt{n}$. If the posterior places almost all its mass in the dashed ellipses, spherical neighbourhoods are inadequate and rates along the axes M and N may differ. More specifically, the rate at which the ellipses decrease in radius may be $(\log n)^{3/2}/\sqrt{n}$ along the M-axis and $1/\sqrt{n}$ along the N-axis.

A more rigorous analysis of the above requires a local reparametrization of the model $\mathscr{P}$ in terms of two new parameters (ϕ, G) in $\mathbb{R}^k \times D$, where the map $(\phi(\theta, F), G(\theta, F))$ is assumed to be continuous, one-to-one and onto in an open neighbourhood U of (θ_0, F_0). Locally around P_0, the model $\mathscr{P}$ can be parametrized equivalently by (θ, F) or (ϕ, G), *i.e.* there exist densities $q_{\phi,G}$ (defined in an open neighbourhood of $(\phi_0, G_0) = (\phi(\theta_0, F_0), G(\theta_0, F_0))$) such that $p_{\theta,F} = q_{\phi(\theta,F),G(\theta,F)}$ for all $(\theta, F) \in U$. For a neighbourhood W of ϕ_0 in Θ and fixed G,

the collection $\{q_{\phi,G} : \phi \in W\}$ defines a parametric submodel $\mathscr{P}_G$ of $\mathscr{P}$. On the submodels $\mathscr{P}_G$, we shall impose regularity conditions, but we try to keep these weak in the sense that they are not required to hold uniform in G, but for each choice of G separately. Furthermore, we impose the requirement that for every submodel $\mathscr{P}_G$, there exists a unique $\phi_G^* \in W$ such that the Kullbach-Leibler divergence with respect to the true distribution P_0 is minimized over W:

$$-P_0 \log \frac{q_{\phi_G^*,G}}{p_0} = \inf_{\phi \in W} -P_0 \log \frac{q_{\phi,G}}{p_0}.$$

Let G_n be a sequence of parameters G defining a sequence $\mathscr{P}_n = \mathscr{P}_{G_n}$ like above, with minima ϕ_n^* for the Kullback-Leibler divergence. Unless G_n happens to be equal to the true G_0, the model $\mathscr{P}_n$ is misspecified. For every n, we also choose a prior Π_n on W and derive a posterior on $\mathscr{P}_n$ based on an *i.i.d.* P_0^n-distributed sample $X_1, \ldots, X_n$. The result is a sequence of parametric models that satisfy the regularity conditions used in chapter 2, if the dependence of the regression functions f_θ on the parameter θ is sufficiently smooth. Moreover, suitable choices of prior exist and testability of P_0 versus alternatives in $\mathscr{P}_n$ at a fixed distance can be related to the problem of testing P_0 versus complements of Hellinger balls in the non-parametric model (which was solved by means of the entropy condition in earlier sections). Hence, based on theorem 2.2, it seems reasonable to assume that the posterior distributions on each of the models $\mathscr{P}_n$ separately converge at rate $1/\sqrt{n}$.

The following lemma proves the intuitively reasonable assertion that convergence at rate $1/\sqrt{n}$ of the posterior measures for a sequence of (misspecified) parametric submodels to their individual Kullback-Leibler minima implies their convergence to the true value of the parameter, if the sequence of minima itself converges at rate $1/\sqrt{n}$. The sequence of submodels may be chosen stochastically.

Lemma 4.18. *Define a stochastic sequence of parametric models $\mathscr{P}_n$ like above. Assume that the sequence of minima ϕ_n^* satisfies:*

$$\sqrt{n}(\phi_n^* - \phi_0) = O_{P_0}(1). \tag{4.41}$$

Furthermore, assume that for each of the (misspecified) models $\mathscr{P}_n$, the posterior concentrates around ϕ_n^* at rate $1/\sqrt{n}$ in P_0-expectation. Then, for every sequence M_n such that $M_n \to \infty$

$$\Pi_n\big(\sqrt{n}\|\phi - \phi_0\| > M_n | X_1, \ldots, X_n\big) \to 0,$$

in P_0-expectation.

Proof Let $\epsilon > 0$ be given. Uniform tightness of the sequence $\sqrt{k}(\phi_k^* - \phi_0)$ implies that there exists a constant K such that:

$$\sup_{k \geq 1} P_0^n\big(\sqrt{k}\|\phi_k^* - \phi_0\| > K\big) < \epsilon. \tag{4.42}$$

Define the events $A_k = \{\sqrt{k}\|\phi_k^* - \phi_0\| \leq K\}$. Let M_n be such that $M_n \to \infty$. Taking $L_{n,k} = M_n/\sqrt{n} + K/\sqrt{k}$, we see that

$$1_{A_k}\Pi_n\Big(\|\phi - \phi_0\| > L_{n,k} \;\Big|\; X_1, \ldots, X_n\Big) \leq \Pi_n\Big(\|\phi - \phi_k^*\| > M_n/\sqrt{n} \;\Big|\; X_1, \ldots, X_n\Big),$$

and therefore,

$$\begin{aligned} &P_0\Pi_n\Big(\|\phi-\phi_0\|>L_{n,k}\;\Big|\;X_1,\ldots,X_n\Big)\\ &\quad\le P_0 1_{A_k}\Pi_n\Big(\|\phi-\phi_0\|>L_{n,k}\;\Big|\;X_1,\ldots,X_n\Big)+P_0(\Omega\setminus A_k)\\ &\quad\le P_0\Pi_n\Big(\|\phi-\phi_k^*\|>M_n/\sqrt{n}\;\Big|\;X_1,\ldots,X_n\Big)+\epsilon. \end{aligned}$$

Let M_n' be given such that $M_n'\to\infty$. Introducing also a strictly positive sequence ϵ_l converging to zero, we see that a corresponding sequence K_l exists such that (4.42) is satisfied for each pair (ϵ_l,K_l). By traversing the sequence ϵ_l slowly enough, we can guarantee that $K_l\le\frac{1}{2}M_l'$ for all $l\ge1$. Replacing the fixed pair (ϵ,K) by the sequence (ϵ_l,K_l) in the above display and specifying to the case $k=n$, we see that the assumed convergence:

$$P_0\Pi_n\Big(\|\phi-\phi_k^*\|>M_n/\sqrt{n}\;\Big|\;X_1,\ldots,X_n\Big)\to0,$$

for every k and all sequences $M_n\to\infty$, implies that

$$P_0\Pi_n\Big(\sqrt{n}\|\phi-\phi_0\|>M_n+K_n\;\Big|\;X_1,\ldots,X_n\Big)\to0.$$

The choice $M_n=\frac{1}{2}M_n'$ then proves the assertion. □

For the misspecified models $\mathscr{P}_n$, $1/\sqrt{n}$-rate of posterior convergence (and also Bernstein-von-Mises-type asymptotic normality) holds if the dependence of the likelihood on the parameter ϕ is sufficiently smooth (see, Kleijn and van der Vaart [58]).

What remains to be shown, is $1/\sqrt{n}$-rate of convergence for the minima ϕ_n^*, given that we still have the freedom to fix the reparametrization and the choice for the model sequence $\mathscr{P}_n$. To see which choices may be expected to lead to the desired result, we turn again to the simplified representation we used earlier, in terms of parameters (θ,F) in $\mathbb{R}^2$. Assuming second-order smoothness of the Kullback-Leibler divergence at P_0, a Taylor expansion in (θ,F) gives rise to the following:

$$\begin{aligned} -P_0\log\frac{p_{\theta,F}}{p_0}&=\tfrac{1}{2}(\theta-\theta_0\,F-F_0)\begin{pmatrix} I_{\theta\theta} & I_{\theta F}\\ I_{F\theta} & I_{FF}\end{pmatrix}\begin{pmatrix}\theta-\theta_0\\ F-F_0\end{pmatrix}+\ldots.\\ &=\tfrac{1}{2}I_{\theta\theta}(\theta-\theta_0)^2+\tfrac{1}{2}I_{FF}(F-F_0)^2+(\theta-\theta_0)I_{\theta F}(F-F_0)+\ldots, \end{aligned}\tag{4.43}$$

up to third-order terms. Figure 4.2 gives the local behaviour of the Kullback-Leibler divergence in terms of level sets, represented as ellipses to reflect the second-order expansion.

The reparametrization we choose is dictated by the axis M': if we choose the coordinate $\phi(\theta,F)$ such that it is (approximately) constant along M', we may hope to double the order of dependence of $\phi_F^*-\phi_0$ on $F-F_0$. The simplest reparametrization that achieves this, is an F-dependent shift D of θ:

$$\begin{cases}\phi(\theta,F)=\theta-D(F),\\ G(\theta,F)=F,\end{cases}\tag{4.44}$$

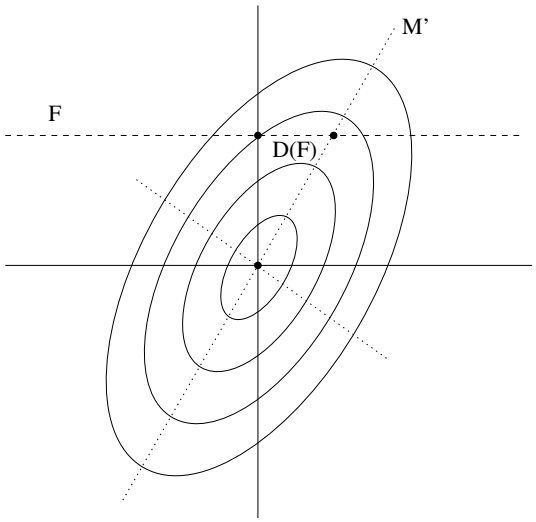

FIGURE 4.2 Local behaviour of the Kullback-Leibler divergence around P_0. The horizontal axis represents the parametric part $\theta \in \Theta$; the vertical axis the non-parametric $F \in D$. The ellipses are level-sets of the Kullback-Leibler-divergence. Note that for every fixed F, the Kullback-Leibler divergence as a function of θ is locally parabolic with its minimum on the axis M'. The difference $D(F)$ is used to shift θ such that the resulting coordinate is constant to second order along M', thus aligning Kullback-Leibler minima.

as indicated in figure 4.2. In that case the models $\mathscr{P}_G$ are simply shifted versions of Θ (locally around θ_0). Substituting (4.44) into (4.43) and requiring the term linear in $F - F_0$ to vanish, we find that with the choice:

$$D(F) = I_{\theta\theta}^{-1} I_{\theta F}(F - F_0),$$

(where it is of course required that $I_{\theta\theta}$ is strictly positive) the same second-order expansion in terms of (ϕ, F) takes the (block-)diagonal form:

$$-P_0 \log \frac{q_{\phi,F}}{p_0} = \tfrac{1}{2}(\phi - \phi_0\, F - F_0) \begin{pmatrix} I_{\theta\theta} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \phi - \phi_0 \\ F - F_0 \end{pmatrix} + \ldots,$$

where the I equals $I_{FF} - I_{\theta F} I_{\theta\theta}^{-1} I_{\theta F}$. Note that $\Delta(F_0) = 0$, so that $\phi_0 = \theta_0$. The upshot is that minimizers of the Kullback-Leibler divergence are all equal to ϕ_0 (*i.e.* are constant along M') up to the order of the above expansion and, in particular, they depend on $F - F_0$ in second order, rather than linearly. To conclude, we note that a consistent maximum-likelihood estimator would approach (θ_0, F_0) along approximately the same axis M', since it minimizes the empirical version of the Kullback-Leibler divergence. In fact, the construction we are describing here coincides with a concept that plays a central role in the semi-parametric

literature and goes under the name of the *least-favourable* direction: it is least-favourable among one-dimensional submodels in the sense that it minimizes the Fisher-information and, as such, provides lower bounds of asymptotic performance for semi-parametric estimation procedures (see, *e.g.* chapter 25 in Van der Vaart (1998) [91]).

In the non-parametric case, we proceed along similar lines. Let $\hat{F}_n = \hat{F}_n(X_1, \ldots, X_n)$ be a sequence of estimators for F_0, converging at rate ϵ_n. This sequence may be chosen as follows: according to previous sections, the posterior concentrates its mass around (θ_0, F_0) at rate $\epsilon_n = (\log n)^{3/2}/\sqrt{n}$. Let $(\hat{\theta}_n, \hat{F}_n)$ be a sequence of (near-)maximisers of the maps:

$$(\theta, F) \mapsto \Pi_n(\, B(\theta, F, \epsilon_n) \,|\, X_1, \ldots, X_n \,),$$

where $B(\theta, F, \epsilon) \subset \Theta \times D$ is the $L_1(\mu)$-ball around (θ, F) of radius ϵ. Such a sequence converges to (θ_0, F_0) at rate $\epsilon_n = (\log n)^{3/2}/\sqrt{n}$, because the posterior is a probability measure concentrating its mass at the required rate. The marginal estimator sequence $\hat{F}_n$ will suffice for our purposes, but other point-estimators converging at comparable rates will suffice for our purposes as well.

Assume that the Kullback-Leibler divergence is finite and twice continuously differentiable in θ in a neighbourhood of (θ_0, F_0). For every fixed F, there is a second-order expansion of the form:

$$-P_0 \log \frac{p_{\theta,F}}{p_0} = \tfrac{1}{2}(\theta - \theta^*_F) V_F (\theta - \theta^*_F) + O(\|\theta - \theta^*_F\|^3), \tag{4.45}$$

where θ^*_F denotes the point at which the Kullback-Leibler divergence is minimal. The first-order term vanishes, since by definition of θ^*_F:

$$\frac{\partial}{\partial \theta} P_0 \log \frac{p_{\theta,F}}{p_0} \,\Big|_{\theta=\theta^*_F} = 0.$$

If $F = F_0$, this equation holds at $\theta = \theta_0$. Subtracting the derivative of $P_0 \log \frac{p_{\theta,F}}{p_0}$ at $\theta = \theta_0$, we obtain:

$$\frac{\partial}{\partial \theta} P_0 \log \frac{p_{\theta,F}}{p_0} \,\Big|_{\theta=\theta^*_F} - \frac{\partial}{\partial \theta} P_0 \log \frac{p_{\theta,F}}{p_0} \,\Big|_{\theta=\theta_0} = \frac{\partial}{\partial \theta} P_0 \Big(\log \frac{p_{\theta,F_0}}{p_0} - \log \frac{p_{\theta,F}}{p_0} \Big) \,\Big|_{\theta=\theta_0}.$$

When we substitute the expansion of the θ-derivative analogous to (4.45), we get:

$$V_F(\theta_0 - \theta^*_F) + O(\|\theta_0 - \theta^*_F\|^2) = \frac{\partial}{\partial \theta} P_0 \Big(\log \frac{p_{\theta,F_0}}{p_0} - \log \frac{p_{\theta,F}}{p_0} \Big) \,\Big|_{\theta=\theta_0}.$$

Assuming that for every $F \in D$, the dependence $\theta \mapsto \log p_{\theta,F}(X)$ is differentiable with respect to θ, P_0-almost-surely, with score $\dot{\ell}_{\theta,F}$, and that differentiation and expectation may be exchanged, we rewrite the above display in the form:

$$V_F(\theta^*_F - \theta_0) + O(\|\theta^*_F - \theta_0\|^2) = P_0\big(\dot{\ell}_{\theta_0,F} - \dot{\ell}_{\theta_0,F_0}\big).$$

Under the assumption that the matrix V_F is invertible, we reparametrize by:

$$\begin{cases} \phi(\theta, F) = \theta - V_F^{-1} \, P_0\big(\Pi_{\theta_0,F} \dot{\ell}_{\theta_0,F}\big), \\ G(\theta, F) = F, \end{cases} \tag{4.46}$$

where the projection $\Pi_{\theta_0,F}$ pertains to the ($L_2(P_0)$-closure of) the tangent space for the nuisance parameter F at the point (θ_0, F). Substituting, we find that up to higher order terms (which can be expressed in different ways, for example with the help of the inverse function theorem), the difference of ϕ_F^* and ϕ_0 behaves like:

$$\phi_F^* - \phi_0 \sim V_F^{-1} P_0\big(\tilde{\ell}_{\theta_0,F} - \tilde{\ell}_{\theta_0,F_0}\big),$$

where $\tilde{\ell}_{\theta,F}$ is the efficient score function for θ. Note that $\phi_F^* = \phi(\theta_F^*, F)$ by the assumption that the minimum of the Kullback-Leibler divergence is unique in small neighbourhoods of the point of convergence. It seems reasonable to expect the *r.h.s.* to satisfy:

$$\big\|P_0\big(\tilde{\ell}_{\theta_0,F} - \tilde{\ell}_{\theta_0,F_0}\big)\big\| = O\big(\|p_{\theta_0,F} - p_{\theta_0,F_0}\|^\alpha\big), \tag{4.47}$$

for some $\alpha > 1$ (possibly even with $\alpha = 2$). Substitution of the sequence $\hat{F}_n$ chosen above then leads to the desired result, since:

$$\big\|p_{\theta_0,\hat{F}_n} - p_{\theta_0,F_0}\big\|_{1,\mu}^\alpha \leq \big\|p_{\hat{\theta}_n,\hat{F}_n} - p_{\theta_0,F_0}\big\|_{1,\mu}^\alpha = O_{P_0}\Big(\frac{(\log n)^{3\alpha/2}}{n^{\alpha/2}}\Big) = O_{P_0}\Big(\frac{1}{\sqrt{n}}\Big).$$

Actual implementation of the above in the errors-in-variables model requires that we shift the parameter θ by an estimator sequence $\Delta_n(X_1, \ldots, X_n)$ for the difference between score and efficient score:

$$\sqrt{n}\Big(\Delta_n - \Delta(\hat{F}_n)\Big) = O_{P_0}(1).$$

Such estimators are known to exist and can be constructed explicitly (see [13]).

A proof of the Bernstein-Von-Mises theorem differs from the above, frequentist plug-in approach in two important respects, because it is based on a marginal posterior distribution for θ: first of all, a major part of the construction shown here stays 'internal' to the proof of a Bernstein-Von-Mises theorem. For example the explicit construction of estimators Δ_n like in the above display is not necessary, because these do not have a place in the conditions or assertions of a Bernstein-Von-Mises theorem. Secondly, we cannot rely on the Kullback-Leibler divergence and its minimizers to 'guide' finite-dimensional Bernstein-Von-Mises assertions along the least-favourable approach to (θ_0, F_0). Instead, the proof would incorporate an expansion of the likelihood (combined with the Jacobian of the reparametrization used in the above) in terms of the efficient score function, expressing local asymptotic normality in a semi-parametric context. Restriction to suitable neighbourhoods of the point of convergence should again follow from sequences of test: one sequence of tests that allow one to restrict to (non-parametric) neighbourhoods converging at the slow, non-parametric rate, and the other, based on the efficient score, allowing one to restrict to parametric neighbourhoods of θ_0 at rate $1/\sqrt{n}$. The latter remarks, however, belong in future work rather than this thesis.

Bibliography

[1] S. Amari, *Differential-geometrical methods in statistics,* Lecture Notes in Statistics No. 28, Springer Verlag, Berlin (1990).

[2] T. Anderson, *Estimating linear statistical relationships,* Ann. Statist. **12** (1984), 1–45.

[3] T. Bayes, *An essay towards solving a problem in the doctrine of chances,* Phil. Trans. Roy. Soc. **53** (1763), 370–418.

[4] S. Bernstein, *Theory of probability,* (in Russian), Moskow (1917).

[5] A. Barron, M. Schervish and L. Wasserman, *The consistency of posterior distributions in nonparametric problems,* Ann. Statist. **27** (1999), 536–561.

[6] J. Berger, *Statistical decision theory and Bayesian analysis,* Springer, New York (1985).

[7] J. Berger and J. Bernardo, *On the development of reference priors,* Bayesian Statistics **4** (1992), 35–60.

[8] R. Berk, *Limiting behaviour of posterior distributions when the model is incorrect,* Ann. Math. Statist. **37** (1966), 51–58.

[9] R. Berk, *Consistency of a posteriori,* Ann. Math. Statist. **41** (1970), 894–906.

[10] R. Berk and I. Savage, *Dirichlet processes produce discrete measures: an elementary proof,* Contributions to statistics, Reidel, Dordrecht (1979), 25–31.

[11] J. Bernardo, *Reference posterior distributions for Bayesian inference,* J. Roy. Statist. Soc. **B41** (1979), 113–147.

[12] P. Bickel and J. Yahav, *Some contributions to the asymptotic theory of Bayes solutions,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **11** (1969), 257–276.

[13] P. Bickel and Y. Ritov, *Efficient estimation in the errors in variables model.* Ann. Statist. **15** (1987), 513–540.

[14] P. BICKEL, Y. RITOV, C. KLAASSEN and J. WELLNER, *Efficient and adaptive estimation for semiparametric models (2nd edition),* Springer, New York (1998).

[15] L. BIRGÉ, *Approximation dans les espaces métriques et théorie de l'estimation,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **65** (1983), 181–238.

[16] L. BIRGÉ, *Sur un théorème de minimax et son application aux tests,* Probability and Mathematical Statistics **3** (1984), 259–282.

[17] L. BIRGÉ and P. MASSART, *From model selection to adaptive estimation,* Festschrift for Lucien Le Cam, Springer, New York (1997), 55–87.

[18] L. BIRGÉ and P. MASSART, *Gaussian model selection*, J. Eur. Math. Soc. **3** (2001), 203–268.

[19] D. BLACKWELL and L. DUBINS, *Merging of opinions with increasing information,* Ann. Math. Statist. **33** (1962), 882–886.

[20] O. BUNKE and X. MILHAUD, *Asymptotic behavior of Bayes estimates under possibly incorrect models,* Ann. Statist. **26** (1998), 617–644.

[21] D. COX, *An analysis of Bayesian inference for non-parametric regression,* Ann. Statist. **21** (1993), 903–924.

[22] H. CRAMÉR, *Mathematical methods of statistics,* Princeton University Press, Princeton (1946).

[23] A. DAWID, *On the limiting normality of posterior distribution,* Proc. Canad. Phil. Soc. **B67** (1970), 625-633.

[24] P. DIACONIS and D. FREEDMAN, *On the consistency of Bayes estimates,* Ann. Statist. **14** (1986), 1–26.

[25] P. DIACONIS and D. FREEDMAN, *On inconsistent Bayes estimates of location,* Ann. Statist. **14** (1986), 68–87.

[26] P. DIACONIS and D. FREEDMAN, *Consistency of Bayes estimates for nonparametric regression: Normal theory,* Bernoulli, **4** (1998), 411-444.

[27] J. DOOB, *Applications of the theory of martingales,* Le calcul des Probabilités et ses Applications, Colloques Internationales du CNRS, Paris (1948), 22–28.

[28] R. DUDLEY, *Real analysis and probability,* Wadsworth & Brooks-Cole, Belmont (1989).

[29] B. EFRON, *Defining curvature on a statistical model,* Ann. Statist. **3** (1975), 1189–1242.

[30] M. ESCOBAR and M. WEST, *Bayesian density estimation and inference with mixtures,* Journal of the American Statistical Association **90** (1995), 577–588.

[31] J. Fan and Y. Truong, *Nonparametric regression with errors in variables,* Ann. Statist. **21** (1993), 1900–1925.

[32] T. Ferguson, *A Bayesian analysis of some non-parametric problems,* Ann. Statist. **1** (1973), 209–230.

[33] T. Ferguson, *Prior distribution on the spaces of probability measures,* Ann. Statist. **2** (1974), 615–629.

[34] D. Freedman, *On the asymptotic behavior of Bayes estimates in the discrete case I,* Ann. Math. Statist. **34** (1963), 1386–1403.

[35] D. Freedman, *On the asymptotic behavior of Bayes estimates in the discrete case II,* Ann. Math. Statist. **36** (1965), 454–456.

[36] D. Freedman, *On the Bernstein-von Mises theorem with infinite dimensional parameters,* Ann. Statist. **27** (1999), 1119–1140.

[37] S. van de Geer, Empirical Processes in M-Estimation. Cambridge University Press, Cambridge (2000).

[38] S. Ghosal, J. Ghosh and R. Ramamoorthi, *Non-informative priors via sieves and packing numbers,* Advances in Statistical Decision theory and Applications (eds. S. Panchapakeshan, N. Balakrishnan), Birkhäuser, Boston (1997).

[39] S. Ghosal, J. Ghosh and A. van der Vaart, *Convergence rates of posterior distributions,* Ann. Statist. **28** (2000), 500–531.

[40] S. Ghosal and A. van der Vaart, *Rates of convergence for Bayes and Maximum Likelihood estimation for mixtures of normal densities,* Ann. Statist. **29** (2001), 1233–1263.

[41] J. Ghosh and R. Ramamoorthi, *Bayesian Nonparametrics,* Springer Verlag, Berlin (2003).

[42] P. Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika **82** (1995), 711–732.

[43] J. Hájek, *A characterization of limiting distributions of regular estimates*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **14** (1970), 323–330.

[44] J. Hájek, *Local asymptotic minimax and admissibility in estimation*, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability **1**, 175–194. University of California Press, Berkeley (1972).

[45] T.-M. Huang, *Convergence rates for posterior distributions and adaptive estimation,* Carnegie Mellon University, preprint (accepted for publication in Ann. Statist.).

[46] P. HUBER, *The behavior of maximum likelihood estimates under nonstandard conditions,* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability **1**, 221–233. University of California Press, Berkeley (1967).

[47] I. IBRAGIMOV and R. HAS'MINSKII, *Statistical estimation: asymptotic theory,* Springer, New York (1981).

[48] D. JACKSON, *The theory of approximation,* American Mathematical Society Colloquium Publications, Vol. XI, New York (1930).

[49] W. JAMES and C. STEIN, *Estimation with quadratic loss,* Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability **1**, 311–319. University of California Press, Berkeley (1961)

[50] H. JEFFREYS, *An invariant form for the prior probability in estimation problems,* Proc. Roy. Soc. London **A186** (1946), 453–461.

[51] H. JEFFREYS, *Theory of probability (3rd edition),* Oxford University Press, Oxford (1961).

[52] R. KASS and A. RAFTERY, *Bayes factors,* Journal of the American Statistical Association **90** (1995), 773–795.

[53] R. KASS and L. WASSERMAN, *A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion,* Journal of the American Statistical Association **90** (1995), 928–934.

[54] M. KENDALL and A. STUART, *The advanced theory of statistics, Vol. 2, (4th edition),* Griffin, London (1979).

[55] YONGDAI KIM and JAEYONG LEE, *The Bernstein-von Mises theorem of survival models,* (accepted for publication in Ann. Statist.)

[56] YONGDAI KIM and JAEYONG LEE, *The Bernstein-von Mises theorem of semiparametric Bayesian models for survival data,* (accepted for publication in Ann. Statist.)

[57] B. KLEIJN and A. VAN DER VAART, *Misspecification in Infinite-Dimensional Bayesian Statistics.* (accepted for publication in Ann. Statist., see chapter 3 in this thesis).

[58] B. KLEIJN and A. VAN DER VAART, *The Bernstein-Von-Mises theorem under misspecification.* (to be submitted for publication in Ann. Statist., see chapter 2 in this thesis).

[59] B. KLEIJN and A. VAN DER VAART, *A Bayesian analysis of errors-in-variables regression,* (to be submitted for publication in Ann. Statist., see chapter 4 in this thesis).

[60] A. Kolmogorov and V. Tikhomirov, *Epsilon-entropy and epsilon-capacity of sets in function spaces,* American Mathematical Society Translations (series 2), **17** (1961), 277–364.

[61] P. Laplace, *Mémoire sur la probabilité des causes par les évenements,* Mem. Acad. R. Sci. Presentés par Divers Savans **6** (1774), 621–656. (Translated in Statist. Sci. **1**, 359–378.)

[62] P. Laplace, *Théorie Analytique des Probabilités (3rd edition),* Courcier, Paris (1820).

[63] L. Le Cam, *On some asymptotic properties of maximum-likelihood estimates and related Bayes estimates,* University of California Publications in Statistics, **1** (1953), 277–330.

[64] L. Le Cam, *On the assumptions used to prove asymptotic normality of maximum likelihood estimators,* Ann. Math. Statist. **41** (1970), 802–828.

[65] L. Le Cam, *Limits of Experiments,* Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability **1**, 245–261. University of California Press, Berkeley (1972).

[66] L. Le Cam, *Convergence of estimates under dimensionality restrictions,* Ann. Statist. **22** (1973), 38–55.

[67] L. Le Cam, *Asymptotic methods in statistical decision theory,* Springer, New York (1986).

[68] L. Le Cam and G. Yang, *Asymptotics in Statistics: some basic concepts,* Springer, New York (1990).

[69] G. Lorentz, *Approximation of functions,* (2nd edition), Chelsea Publ. Co., New York (1986).

[70] R. Megginson, *An introduction to Banach Space Theory,* Springer, New York (1998).

[71] R. Von Mises, *Wahrscheinlichkeitsrechnung,* Springer Verlag, Berlin (1931).

[72] J. Munkres, *Topology (2nd edition),* Prentice Hall, Upper Saddle River (2000).

[73] J. Pfanzagl, *Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures,* J. Statist. Planning Inference **19** (1988), 137–158.

[74] D. Pollard, *Convergence of Stochastic Processes,* Springer, New York (1984).

[75] D. Pollard, *Another look at differentiability in quadratic mean,* Festschrift for Lucien Le Cam, Springer, New York (1997), 305–314.

[76] D. Pollard, *Lecture Notes on Le Cam theory (chapter 7)*, Lectures given in the Odyssey program, Borel Center (IHP), Paris (March-May 2001).

[77] C. RAO, *Information and the accuracy attainable in the estimation of statistical parameters,* Bull. Calcutta Math. Soc. **37** (1945), 81–91.

[78] C. ROBERT, *The Bayesian choice: from decision-theoretic foundations to computational implementation,* Springer, New York (2001).

[79] B. RIPLEY, *Pattern recognition and neural networks*, Cambridge University Press, Cambridge (1996).

[80] O. REIERSØL, *Identifiability of a linear relation between variables which are subject to error,* Econometrica **18** (1950), 375–389.

[81] M. SCHERVISH, *Theory of statistics,* Springer, New York (1995).

[82] L. SCHWARTZ, *On Bayes procedures,* Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **4** (1965), 10–26.

[83] X. SHEN and L. WASSERMAN, *Rates of convergence of posterior distributions,* Ann. Statist. **29** (2001), 687–714.

[84] X. SHEN, *Asymptotic normality of semiparametric and nonparametric posterior distributions,* Journal of the American Statistical Association **97** (2002), 222–235.

[85] H. STRASSER, *Mathematical Theory of Statistics,* de Gruyter, Amsterdam (1985).

[86] M. TAUPIN, *Semi-parametric estimation in the nonlinear structural errors-in-variables model,* Ann. Statist. **29** (2001), 66–93.

[87] A. VAN DER VAART, *Estimating a real parameter in a class of semiparametric models,* Ann. Statist. **16** (1988), 1450–1474.

[88] A. VAN DER VAART, *Efficient maximum-likelihood estimation in semiparametric mixture models,* Ann. Statist. **24** (1996), 862–878.

[89] A. VAN DER VAART and J. WELLNER, *Weak Convergence and Empirical Processes,* Springer, New York (1996).

[90] A. VAN DER VAART, *Superefficiency*, Festschrift for Lucien Le Cam, Springer, New York (1997), 397–410.

[91] A. VAN DER VAART, *Asymptotic Statistics,* Cambridge University Press, Cambridge (1998).

[92] A. VAN DER VAART, *Semiparametric statistics,* Lectures on Probability Theory, Ecole d'Eté de probabilités de St. Flour (ed. P. Bernard), Springer Verlag, Berlin (2000).

[93] A. VAN DER VAART, *Limits of Experiments*, (to be published by Cambridge University Press).

[94] A. Wald, *Note on the consistency of the maximum likelihood estimate*, Ann. Math. Statist. **20** (1949), 595–601.

[95] A. Walker, *On the asymptotic behaviour of posterior distributions,* J. Roy. Statist. Soc. **B31** (1969), 80–88.

[96] L. Wasserman, *Bayesian model selection and model averaging*, J. Math. Psych. **44** (2000), 92–107.

[97] W. Wong and X. Shen, *Probability inequalities for likelihood ratios and convergence rates of sieve MLE's, Ann. Statist.* **23** (1995), 339–362.

[98] Y. Yang and A. Barron, *An asymptotic property of model selection criteria*, IEEE Transactions on Information Theory **44** (1998), 95–116.

Samenvatting

Een statistisch model is misgespecificeerd indien het de werkelijke verdeling van de data niet bevat. Alleen in het geval dat het model volledig is, in de zin dat het alle verdelingen bestrijkt die de data mogelijkerwijze kan hebben, kan men deze situatie op voorhand uitsluiten. De reden voor het gebruik van misgespecificeerde modellen is niet alleen het feit dat volledige modellen in praktijk te groot zijn voor toepassing van geëigende stellingen en schattingsmethodes, maar ook vanwege de interpretatie die verleend kan worden aan de parameters in kleinere modellen. Niettemin wordt op grote schaal gebruik gemaakt van de aanname dat het model goed gespecificeerd is, zelfs in situaties waarin dit zeer onwaarschijnlijk geacht kan worden. Het feit dat die veronderstelling dikwijls zonder consequenties blijft, doet vermoeden dat de voorwaarde van een goed gespecificeerd model in stellingen tot op zekere hoogte overbodig is, of vervangen kan worden door zwakkere voorwaardes op het model en de verdeling van de data. Dit proefschrift kan worden samengevat als een poging dergelijke compatibiliteitsvoorwaardes voor model en data te vinden, waar het stellingen in de asymptotiek van Bayesiaanse methodes betreft.

Hoofdstuk 1 heeft tot doel de meest belangrijke concepten te introduceren op een zo eenvoudig mogelijk niveau, als inleiding op latere hoofdstukken waarin dezelfde concepten in ingewikkeldere vorm terugkomen. Nadrukkelijk is getracht, analogieën tussen puntschatting en Bayesiaanse methodes naar voren te brengen, voornamelijk om de toegankelijkheid voor een breed publiek te garanderen. Met uitzondering van de laatste sectie, gaat het gehele eerste hoofdstuk uit van een goed gespecificeerd model. Na een zeer beknopte introductie aangaande de keuze van een prior en de definitie van de posterior, wordt het asymptotisch gedrag van puntschatters besproken. Een en ander wordt geïllustreerd aan de hand van de maximum-likelihood schatter in parametrische modellen. Tevens wordt ingegaan op lokaal-asymptotisch-normaal gedrag van de likelihood en asymptotische optimaliteitscriteria voor puntschatters. Vervolgens concentreert zich de discussie op stellingen aangaande consistentie (Doob, Schwarz), convergentiesnelheid (Ghosal–Ghosh–van-der-Vaart) en limietvorm (Bernstein–Von-Mises) van de posterior. Tevens wordt toegelicht welke invloed dergelijke eigenschappen van de posterior hebben op Bayesiaanse puntschatters. Het hoofdstuk sluit af met een voorbeschouwing van de invloed van misspecificatie op de argumentatie.

Hoofdstuk 2 behandelt het limietgedrag van een posterior verdeling voor misgespeci-

ficeerde, gladde, parametrische modellen. Het Bernstein–Von-mises theorema stelt dat, in goed gespecificeerde modellen en onder bepaalde gladheidsvoorwaardes, de posterior convergeert naar een normale verdeling. De laatste is gelocaliseerd op de maximum-likelihood schatter en heeft als covariantie de inverse Fisher-informatie gedeeld door de grootte van het sample. In een misgespecificeerd model kan een vergelijkbare stelling worden bewezen, waarbij het punt van convergentie de Kullback-Leibler projectie van de ware verdeling is en de snelheid van convergentie parametrisch blijft. De prijs die betaald wordt voor misspecificatie presenteert zich in de vorm van een grotere asymptotische variantie. Ook wordt aandacht besteed aan stellingen aangaande convergentie-snelheid en consistentie onder zwakkere voorwaarden.

Hoofdstuk 3 behandelt de convergentiesnelheid van de posterior in niet-parametrische modellen. De oneindige dimensie van het model heeft uiteenlopende consequenties: om te beginnen is de definitie van een prior die zijn massa homogeen verdeelt over een oneindig-dimensionaal model niet triviaal. Voorts is het bestaan van bruikbare rijen van test-functies van kritiek belang. Beide problemen doen zich ook voor in het goed gespecificeerde geval. Priors op modellen van waarschijnlijkheidsmaten stonden reeds in de jaren '60–'70 volop in de belangstelling. Het bestaan van test-functies kan worden afgeleid uit het gedrag van de (Hellinger) metrische entropie van het model. In misgespecificeerde context blijkt de constructie van bruikbare priors analoog, maar moet voor het bestaan van test-rijen aanzienlijk meer werk verricht worden. De convergentiesnelheid blijkt echter in veel gevallen niet gevoelig voor model misspecificatie.

Hoofdstuk 4 geeft een Bayesiaanse analyse van Errors-In-Variables regressie. Het model beschrijft twee variabelen waartussen bij benadering een functioneel verband bestaat. Niet alleen is er sprake van "ruis" in beide gemeten grootheden, maar bovendien hangen beiden af van een derde, ongemeten grootheid met onbekende verdeling. Het doel is schatting van het functioneel verband. In de semi-parametrische literatuur wordt een parametrische familie van regressie functies beschouwd en wordt de onbekende verdeling van de derde grootheid als niet-parametrische nuisance behandeld. In het model dat in dit hoofdstuk wordt beschreven, wordt ook de regressie-familie oneindig-dimensionaal verondersteld. Convergentiesnelheid voor de posterior wordt gekwantificeerd in termen van de Hellinger afstand tussen dichtheden voor het paar van gemeten grootheden. Het model heeft in dat geval twee niet-parametrische componenten. Het blijkt dat de convergentiesnelheid van de posterior bepaald wordt door de langzaamst convergerende van de twee. Ook in dit geval zijn metrische entropy van het model en uniformiteit van de prior bepalend voor het asymptotisch gedrag. De semi-parametrische analyse wordt gemaakt in het tweede, deels speculatieve gedeelte van dit hoofdstuk: indien de familie van regressie functies parametrisch gekozen wordt, dient zich de mogelijkheid aan om het materiaal uit hoofdstuk 2 te gebruiken in een bewijs van parametrische convergentiesnelheid. De methode is gebaseerd op de zogenaamde "least-favourable direction", die ook centraal staat in de semi-parametrisch benadering van het probleem. Een en ander biedt tevens uitzicht op een semi-parametrische versie van het Bernstein–Von-Mises bewijs.

Dankwoord

Eerst en vooral gaat mijn dank uit naar mijn promotor Aad van der Vaart, voor de mogelijkheid die hij mij geboden heeft om een promotieonderzoek binnen de mathematische statistiek te verrichten en de begeleiding die hij daarbij gegeven heeft. De discussies met betrekking tot het onderzoek, zijn commentaar op stukken werk en zijn visie op de hedendaagse statistiek blijken een uitstekende voorbereiding te vormen op mijn huidige werk in Berkeley.

Verder bedank ik mijn collega's aan de Vrije Universiteit voor alle plezierige lunches en talloze interessante discussies. Met name noem ik mijn beide kamergenoten, Martine Reurings en Frank van der Meulen, aan wiens gezelschap ik zeer prettige herinneringen overhoud.

Met betrekking tot het proefschrift zelf, bedank ik allereerst Aad voor het lezen en becommentariëren van talrijke versies van het manuscript. Daarnaast ben ik ook de leescommissie, bestaande uit Peter Bickel, Sara van de Geer, Richard Gill en Piet Groeneboom erkentelijk voor de tijd die zij hebben geïnvesteerd in het lezen van de uiteindelijke tekst en het commentaar dat zij gegeven hebben. I thank Peter in addition for numerous illuminating discussions we have had in Berkeley and his attendance at my thesis-defence.

Verder bedank ik mijn vrienden Eric Cator en Baldus Tieman voor de moeite die zij, in mijn afwezigheid, gedaan hebben in de voorbereiding van de promotie (met name in het drukken van het proefschrift) en natuurlijk voor het feit dat zij (beiden voor de tweede keer) willen optreden als mijn paranimfen. Tenslotte bedank ik mijn familie en mijn vrouw Suzan, zonder wiens steun de voltooiing van het onderzoek en in het bijzonder dit proefschrift een veel grotere opgave zouden zijn geweest.

Curriculum Vitae

Ik ben geboren op 15 oktober 1970, te Nijmegen. Van 1983 tot 1989 bezocht ik het Rythovius College in Eersel, alwaar ik in juni 1989 het eindexamen VWO behaalde. Vervolgens ben ik begonnen met de studies Wiskunde en Natuurkunde aan de Universiteit Utrecht. In juni 1990 ontving ik de propaedeuse Natuurkunde en de propaedeuse Wiskunde (cum laude) en in augustus 1994 volgde het doctoraal examen Natuurkunde (cum laude). Mijn doctoraalscriptie theoretische natuurkunde heb ik geschreven onder begeleiding van Prof. dr. B.Q.P.J. de Wit, die ook de supervisie van mijn eerste promotieonderzoek op zich heeft genomen. Dat onderzoek, gestart in september 1994, verricht aan het instituut voor theoretisch fysica te Utrecht in dienst van de stichting FOM, leidde in mei 1998 tot promotie op een proefschrift getiteld *New couplings in $N = 2$ supergravity.*

Vervolgens diende zich een moeilijke keuze aan: de sombere carrièreperspectieven binnen de theoretische hoge-energie fysica en de vlucht die de natuurkundige verbeelding in mijn ogen genomen had in de vorm van de 'tweede string-revolutie', boden een weinig aanlokkelijk toekomstbeeld. Feit bleef dat de theoretische natuurkunde een prachtig vakgebied was, waarin ik altijd met veel plezier gewerkt had. Uiteindelijk besloot ik mijn heil desondanks elders te zoeken: aanvankelijk in de medische beeldverwerking, in de vorm van een éénjarige postdoc-aanstelling onder begeleiding van Prof. dr. M.A. Viergever bij het Image Sciences Institute aan de faculteit Geneeskunde van de Universiteit Utrecht en vervolgens in de mathematische statistiek, als promoverend postdoc onder begeleiding van Prof. dr. A.W. van der Vaart bij de afdeling Wiskunde van de Faculteit der Exacte Wetenschappen van de Vrije Universiteit te Amsterdam. Het resultaat van de drie-en-half jaar onderzoek die ik in die laatste hoedanigheid gedaan heb, ligt voor u.

Sinds 1 september 2003 werk ik als postdoc aan het Statistics Department van U.C. Berkeley onder begeleiding van Prof. dr. P.J. Bickel, op een zogeheten TALENT-stipendium ter financiering van een jaar verblijf en onderzoek aan een buitenlands instituut, beschikbaar gesteld door de stichting NWO.

Cover Illustrations

The figure on the front cover originates from Bayes (1763), *An essay towards solving a problem in the doctrine of chances,* (see [3] in the bibliography), and depicts what is nowadays known as Bayes' Billiard. To demonstrate the uses of conditional probabilities and Bayes' Rule, Bayes came up with the following example: one white ball and n red balls are placed on a billiard table of length normalized to 1, at independent, uniformly distributed positions. Conditional on the distance X of the white ball to one end of the table, the probability of finding exactly k of the n red balls closer to that end, is easily seen to be:

$$P\big(\, k \;\big|\; X = x \,\big) = \frac{n!}{k!\,(n-k)!}\, x^k (1-x)^{n-k}.$$

One finds the probability that k red balls are closer than the white, by integrating with respect to the position of the white ball:

$$P(\, k \,) = \frac{1}{n+1}.$$

Application of Bayes' Rule then gives rise to a Beta-distribution $B(k+1, n-k+1)$ for the position of the white ball conditional on the number k of red balls that are closer. The density:

$$\beta_{k+1,n-k+1}(x) = \frac{(n+1)!}{k!\,(n-k)!} x^k (1-x)^{n-k},$$

for this Beta-distribution is the curve drawn at the bottom of the billiard in the illustration.

The illustation on the back of the cover shows (part of) Thomas Bayes' autograph.